nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

FIVE HUBS OF ASIAN SCIENCE

Research and investment in Hong Kong, Malaysia, Singapore, South Korea and Taiwan PAGE 499

STRUCTURAL BIOLOGY

COMPLEX PROCESS

Four structures shed light on G protein selectivity

PAGES 529, 547, 553, 559 & 620

PLANETARY SCIENCE

THE ROCKY ROAD TO MARS

Meteorites draw timeline for formation of the red planet

PAGES 522 & 586

MEDICAL RESEARCH

CAUSE AND EFFECT

The mechanism behind weight loss in cancer

PAGES 526 & 600

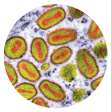
NATURE.COM/NATURE
28 June 2018
Vol. 558, No. 7711

THIS WEEK

EDITORIALS

MOVING ON *Nature*'s editor-in-chief says farewell after more than two decades **p.486**

WORLD VIEW Look to computers to draw fairer electoral districts **p.487**



VIRAL LOAD Protein cage to hold viruses offers new-drug tips **p.489**

Worrying changes in Hungary

The European country's autocratic government has made a disturbing grab at the nation's scientific institutions.

Travel writers like to call Hungary a land of contrasts. But the cliché has been true there in recent years, as the ultranationalist government of Viktor Orbán has tightened its grip. Although European Union politicians have watched Hungary's increasingly anti-democratic tendencies with mounting concern, researchers have seen the nation's research base begin to flourish, with new internationally competitive laboratories.

This juxtaposition has been achieved because, until now, Hungary has left science in the hands of its own experts. And for the most part, they have done a splendid job. That situation has now changed. The authoritarian government is snatching away scientific autonomy — and this should provoke alarm.

The storm has been gathering since an April election brought a landslide victory for Orbán's Fidezs party. In its path is the proud Hungarian Academy of Sciences, which has stood independent of politics for more than two decades, since the collapse of communism in the region. The academy has done sterling work, including creating major research grants that allowed many researchers who had been working abroad to return to Hungary and establish independent labs. Yet the government's budget proposal for next year, announced earlier this month, would transfer the majority of the academy's financing into the newly created Ministry for Innovation and Technology.

And last week, Orbán dismissed József Pálinkás from the leadership of the National Research, Development and Innovation Office, a post he has held for three years. Since 2015, Pálinkás, a physicist and former academy president, has created from scratch a broad portfolio of funding programmes ranging from basic science to near-industry research. The scheme was a model for how to build a science base founded on excellence, and it triggered a welcome reversal to Hungary's previous brain drain (see *Nature* **551**, 425–426; 2017).

According to innovation minister László Palkovics, the changes are to unify innovation and science policy, and to eliminate fragmentation of research budgets. On the face of it, that's reasonable. Palkovics promises that the academy money will be filtered back to its various research institutions. And any incoming government in a democratic country, of course, has the right and mandate to replace key members of staff.

Yet many researchers in Hungary tell *Nature* they are worried academy money might be returned with strings attached — maybe instructions that it should be spent to serve the economy more directly, or that historians should glorify their country's past. Trust is at a low ebb. The government's actions in other areas are becoming ever more extreme. On the day of Pálinkás's dismissal, for example, the parliament approved a law that makes helping refugees to apply for asylum in Hungary a crime punishable by up to 90 days' imprisonment. It also approved constitutional changes that require all state institutions to protect Hungarian cultural and Christian values, and that make homelessness illegal.

Orbán has never felt comfortable with what he sees as academia's international and elitist air. A particular bugbear for him has been the Central European University (CEU), which was founded in 1991 by Hungarian-born philanthropist George Soros, and is registered in New York state but located in Budapest. A law rushed through in April last year required international universities to operate as higher-education institutes in their country of origin as well as in Hungary.

That law affected only the CEU, whose agreement to remain in Hungary expires at the end of this year. The change attracted an impressive 70,000 protesters to the streets of Budapest, and the CEU quickly arranged higher-educational activities in the United States to be compliant with the law. But the government has still not signed off on a new agreement for the university to stay in the country, drafted last September. Negotiations are continuing, but the CEU has organized an alternative home for itself in Austria, and a transfer there seems increasingly likely as deadlines for recruiting next year's students approach. That would deprive Hungary of a valuable intellectual hub, and would mark another significant step backwards for the country.

> ***"The authoritarian government is snatching away scientific autonomy."***

The message to Hungary should be clear: ensure that the government's new management and methods continue to uphold the principles of meritocratic funding. And maintain the possibilities of long-term funding for excellent basic research, to help ensure that a strong scientific community can continue to feed the government's laudable innovation ambitions. Meanwhile, the 2019 budget, with its plan to take control of the Hungarian Academy of Sciences' funds, is scheduled to be approved by mid-July. There is still time for the government to reverse its course. It should do so. ■

Local science

Researchers in five Asian economies are working to address communities' needs.

One of the most commonly stated goals of science and scientists is to work to improve society. But which society? The needs and circumstances of people, communities and regions across the world are very different — from energy use and disease threats to natural-resource availability and pollution.

In a special issue this week, *Nature* explores how some of these local needs are being addressed across five strong science centres in East Asia: Hong Kong, Malaysia, Singapore, South Korea and Taiwan. Over the past few decades, each member of this diverse group has evolved its own model of how to pursue research successfully. Impressively, some of their key achievements are those in which they have matched the science agenda to explicit and unique local requirements.

That is a good model for others to follow, especially given that large numbers of people around the world are not well enough served by the agendas and interests that drive much of modern science. *Nature* has argued before that more scientists and funders should reach out to identify and tackle direct societal challenges in this way (*Nature* **542**, 391; 2017).

Each of the economies we highlight has a unique history that has shaped its research and development. Take Malaysia, a peninsula and constellation of islands sandwiched between Thailand and Indonesia. In the 1970s, it started to shift its economic reliance on cheap products such as tin, rubber and cocoa to higher-value commodities such as natural gas and palm oil. It then used applied science to foster a booming electronics industry. Some of its major exports today include other products of applied research, including chemicals. Yet this success has come with relatively low investment in science and technology.

More unusual still, almost half of researchers there are women (see page 500). In a News Feature this week (see page 502), among others, we profile Malaysian chemical engineer Suzana Yusup, who leads a centre that makes fuels from biomass waste, such as used cooking oil, rubber-seed oil and discarded distillate from palm-oil refineries. Her career has focused on green technologies that can help the environment and society.

Scientists in Malaysia, which has a predominantly Muslim population, are also developing Halal substitute ingredients for food, pharmaceuticals and cosmetics, reaping the benefits of a Halal economy that, in 2016, was worth US$2 trillion globally.

Malaysia demonstrates how applied science can generate the economic benefits that can allow officials to invest in societal needs. And it's not alone. Singapore, along with South Korea and Taiwan, has long focused on applied projects across electronics, physics and materials science. The success of these has boosted its gross domestic product (GDP). And the Singaporean government is now putting some of that money into national priorities — health care and biomedical sciences among them. There's a strong push to understand, detect and treat heart disease and cancers of the liver, stomach, breast and lung, which have a significant impact on Asian populations.

By most measures, South Korea is an impressive performer in science, making it a giant in the region. It invests more than 4% of its GDP in science and technology — much of it applied — and has a high density of researchers per head of population. Its output of scholarly articles has skyrocketed in the past two decades.

"Each of the economies has a unique history that has shaped its research and development."

But South Korea is also choking under a cloud of air pollution, and, as physicist Han Woong Yeom at Pohang University of Science and Technology writes in a Comment piece (see page 511), its science policy must be updated to address this and other national needs. That might already be happening. A 2015 analysis of development of the regional research and technology organization in Gyeonggi province suggested that policymakers had switched from a top-down approach to one that emphasizes the "detailed analysis of local industry needs" (S. Shin *Reg. Stud. Reg. Sci.* **2**, 424–431; 2015).

Not all scientists in conventional research-powerhouse economies might welcome such direct targeting of local problems, just as some frown when politicians talk up the need for applied science. But there does not have to be a trade-off between work that is of international quality and work that has a direct local impact. Hong Kong, for example, has grown into a hub for researchers investigating emerging infectious diseases, such as the avian influenza strain H5N1 and severe acute respiratory syndrome (SARS), both of which originated in Asia. Those teams published papers in leading journals. But research there has also demonstrated that closing live-poultry markets for a day or two each month could dramatically reduce the spread of bird flu and cut the risk to people. That's a win–win situation, the likes of which all societies should encourage — wherever they are. ■

Valediction

A reflection as the seventh editor-in-chief of Nature *hands over to the eighth.*

This issue of *Nature* is the last under my tenure as the publication's editor-in-chief. The first was published on 14 December 1995. A few personal thoughts seem in order.

Nature's editorial role since its foundation in 1869 has consistently been about support for outstanding science while also being a critically minded friend of the research community and its values. Fired by my own enthusiasms for astronomy and physics since childhood and as a researcher, and by this publication's ever-broadening interests and international ethos, it has been my extraordinary good fortune and privilege to work with many researchers and colleagues to help *Nature* to continue and develop in its mission.

As a journal, *Nature* has thrived by keeping abreast of some of the most inspired and inspiring research — insights into the human genome and the microbiome, developments in photovoltaics and the extraordinary flowering of exoplanet research are just some examples that have been a joy to see. The journal has also gratifyingly grown into areas that were well established elsewhere — organic chemistry and high-energy physics are two. And the totally unexpected has always felt best: *Homo floresiensis* ('the Hobbit') was perhaps my own favourite.

On the magazine components, a look back at some 1995 issues shows how focused *Nature* then was on narrow rather than widely interesting policy news, how little commissioned comment there was relating to the research enterprise and its external relationships and how impenetrable some of the language was in our News and Views section. Ever since, it has always been my ambition and that of the editorial teams progressively to open up our pages to more lively and comprehensible fare.

My regrets include wonderful papers that we failed to attract, and that we still have more to do in speeding up our handling of labyrinthine complexities that can arise in retractions and formal critiques of our papers. There are initiatives under way towards being more attentive in our content to the needs and interests of under-represented groups in the population and in the research community, and being equivalently more diverse in the make-up of our editorial team. I wish I had pushed harder on all of these fronts.

An editor-in-chief has a platform on which to champion readers' needs and interests — and also under-attended causes. Mine have included the interests of social sciences, reproducibility, healthy research cultures and environments, the tracking of research's societal impacts, and mental-health research. Throughout, my goal has been, above all, to make the weekly issue — much of it now published continuously online — something that as many as possible of our very demanding audience eagerly look forward to.

Whatever has been achieved, none of it would have been possible without great colleagues. *Nature*'s editorial staff over the past 22-plus years has included many inspiringly skilled and visionary individuals. As a result, while there have been some acknowledged missteps, the time we have spent has been rich in fulfilment — at least for me, possibly for them too, and above all, I hope, for readers.

As I move on to a new role as editor-in-chief of our publishing company Springer Nature, I thank those many people inside and outside the research community who have helped to make *Nature* what it is. Above all, I offer the *Nature* team my profound thanks. I wish them and my successor Magdalena Skipper all the very best in their abundant future responsibilities and opportunities. **Philip Campbell**

BRIAN STAUFFER

Algorithms can foster a more democratic society

Counterbalancing the Supreme Court's gerrymandering ruling is technology's potential to prevent gerrymanders in the first place, says **Wendy K. Tam Cho**.

Gerrymandering — the manipulation of district boundaries to give one group a political advantage — is not part of anyone's idea of democracy. Although it is difficult to define gerrymandering precisely, the contorted shapes of electoral districts defy simple explanation and imbue a public perception of a rigged system. And when, as in Pennsylvania in 2014, a party captures 72% of its US House of Representatives seats with only 55.5% of the statewide vote, suspicions are piqued.

Thus this week's decision by the Supreme Court, which all but squelched hopes for a manageable standard ahead of the 2020 redistricting cycle, is unwelcome news for those who anticipated that the court would take a forceful lead in curtailing partisan gerrymandering. However, even with such a standard for detecting gerrymanders, politicians have shown us that they are extremely savvy when it comes to circumventing legal constraints.

An ounce of prevention is worth a pound of cure. I argue that the means of such prevention lies not with the courts but in technological advances, as long as we are mindful of Supreme Court Justice Anthony Kennedy's admonishment in 2004 that, for partisan gerrymandering, "technology is both a threat and a promise".

In the United States, electoral districts are redrawn every ten years. In more than two-thirds of the states, partisan legislators control congressional redistricting. A proliferation of software that emerged about 30 years ago has facilitated the drawing of electoral maps that simultaneously entrench power while meticulously adhering to legal districting practices. Worse, current redistricting software requires experts with political and legal savvy, who generally work in secret behind closed doors. Hence, the software has served only to advance the threat of technology in redistricting.

We must now work to enable its promise.

I develop statistical and computational models that intelligently extract information. My research uses the world's fastest supercomputers in the service of social progress. For redistricting, this means devising efficient algorithms that make quadrillions of calculations per second on highly sophisticated computing architectures to explore how best to ensure fairness in electoral maps.

The task of redistricting is well suited for computational algorithms because the goals can be articulated clearly, performance metrics can be specified easily and the tasks are distinct and structured. Moreover, computational algorithms are able to present a wide array of possibilities that capture the interests of diverse societal groups. Perhaps most importantly, computers are impervious to the lure of power.

Because our collective voice is composed of the individual voices of many distinct and diverse groups, political fairness is a complex phenomenon. It requires compromise and balancing competing interests so that members of all groups (racial and ethnic minorities, labour unions, all socio-economic levels and so forth) are represented.

Citizens and interest groups can articulate what political fairness means to them, but they lack the legal and political expertise to translate their goals into actual electoral maps, so their voices are easily muted. This is where intelligent computational algorithms can play a part. They can search for possible maps that simultaneously adhere to legal thresholds (for example, compactness, representation of minority groups, percentage of split municipal subdivisions) while fulfilling criteria from partisan groups and non-partisan ones, such as the League of Women Voters, and Common Cause, which promote competitive voting districts, and groups such as the American Civil Liberties Union, whose mission it is to protect the civil rights of all Americans. Algorithms could amalgamate these wide and varied interests to identify electoral maps that are acceptable to a broad swathe of society.

COMPUTERS ARE **IMPERVIOUS** TO THE **LURE** OF **POWER.**

Technological innovation could supply missing information that is highly significant for improving democratic society. Maps that encompass competing interests must be made central to redistricting discussions and deliberation by politicians and independent commissions.

Of course, algorithms can themselves embody bias. Concerns include well-publicized issues around 'predictive-policing' programs (see *Nature* **558**, 357–360; 2018) that aim to determine who is at risk of reoffending; these can unfairly penalize African Americans. In the case of redistricting, however, the algorithms are not making decisions, but fostering more-inclusive conversations. The criteria are supplied by diverse groups with valid competing interests. These maps do not become law in secret, but set the stage for deliberative democracy. Humans are free to reject and modify them as they see fit.

That is why my colleague, Yan Liu, and I have been developing PEAR (Parallel Evolutionary Algorithm for Redistricting), a computational algorithm that integrates Supreme Court mandates and carries out intelligent analysis to identify legally viable maps that satisfy an array of specific goals. (PEAR is tailored for the United States, but the core ideas of exploring redistricting possibilities transfer easily to other locales.) Our hope is to move technological advances in the direction of supplying objective information that empowers the inclusion of diverse societal groups and enhances human deliberation.

So far, technology for redistricting has led only to the exclusion and isolation of power. Moving forward, we must harness the power of technology to ensure democracy. The promise of technology is to augment human capabilities to engage in productive, inclusive and contemplative decision-making about how society is governed. ■

Wendy K. Tam Cho *is a professor of political science, law, statistics and mathematics at the University of Illinois at Urbana-Champaign.*
e-mail: wendycho@illinois.edu

SEVEN DAYS *The news in brief*

FACILITIES

Telescope boost

Spain has joined the effort to build the world's largest radio telescope, which will probe the early Universe. When complete, the Square Kilometre Array (SKA) will have about 2,000 radio dishes in South Africa and up to 1 million in Australia. The project's €674-million (US$786-million) first phase was scaled back last year to cut costs; it will now consist of 194 dishes in South Africa and about 130,000 antennas in Australia, and it is due to begin construction in 2020. The project's governing body, the SKA Organisation, has had ten members, including Sweden, the United Kingdom and the two host nations, since Germany left in 2014. Spain will pay an undisclosed membership fee to join, and it will enter into negotiations about its contribution to phase one. The SKA hopes that recruiting new members will bring it the funds needed to build the full design of the first phase, says SKA director of communications William Garnier.

NATURE'S SALARY SURVEY

If you are a working scientist, or you work in a science-based job, we would love to hear from you in *Nature*'s 2018 salary survey. Every two years, we ask you about your job: what you love, what you dislike, what you have done so far and what you would change. The survey takes around 20 minutes to complete, and we will report on the findings later this year. Your answers will help us to identify global trends in income, benefits and job satisfaction, as well as help you to make more-informed career decisions. Survey entrants will have the chance to enter a prize draw to win a £100 or US$150 Amazon gift card. Take the survey by 16 July at go.nature.com/salarysurvey2018

BETTMAN/GETTY

Gorilla who knew sign language dies

Koko, a gorilla that learnt an adapted version of American Sign Language, died in her sleep on 19 June. She was 46 years old. The announcement came from the California-based non-profit group The Gorilla Foundation, whose co-founder, animal psychologist Francine Patterson, started teaching Koko sign language in 1972 while studying the cognitive abilities of western lowland gorillas (*Gorilla gorilla gorilla*). Koko made headlines not only for her language skills — she reportedly used more than 1,000 signs and understood around 2,000 words of spoken English — but also for adopting kittens, the first of which she named All Ball.

POLICY

Trump makeover

US President Donald Trump released a sweeping plan on 21 June to reorganize — and shrink — the federal government. Oversight of fisheries would move from the National Oceanic and Atmospheric Administration, in the commerce department, to the interior department. The plan also calls for the Food and Drug Administration to be renamed the Federal Drug Administration, and for its responsibility of ensuring food safety to be transferred to the agriculture department. But the proposal, which would need to be approved by Congress, already faces stiff opposition from top Democrats and is unlikely to take effect. Trump's predecessor, Barack Obama, tried and failed to implement his own government reorganization plan in 2012.

Embryo research

New Zealand's guidelines on the use of human embryos in research are unclear and a barrier to infertility studies, according to a survey of the country's researchers published in *The New Zealand Medical Journal*. Although the use of human embryos for research is allowed under New Zealand law, guidelines developed in 2005 refer to the use of only non-viable embryos, effectively banning researchers from using viable human embryos, say the survey's authors. Most of the 20 human-embryo researchers surveyed felt that they were disadvantaged by the lack of specific guidance, with 11 saying they had potential research projects that they couldn't take up under the current regulations.

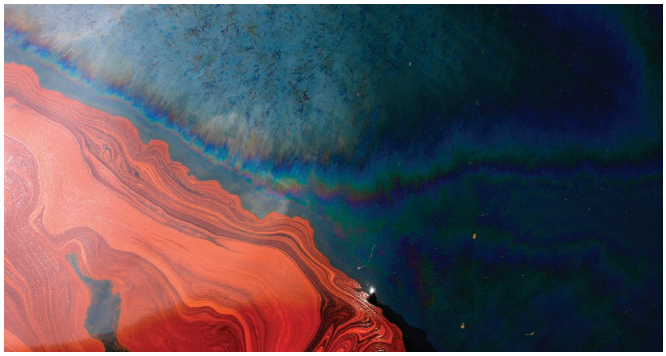
EU copyright laws

A set of controversial changes to copyright rules in the European Union has passed a hurdle on the way to becoming law. The proposed legislation, which covers copyright of

digital material in the EU, would require social-media platforms to pay fees to rightholders for any content or snippets of content uploaded on these platforms. It would also require web services, including repositories for research articles, to actively prevent uploads of copyrighted material. The legislation exempts academic researchers from paying fees to perform data and text-mining — computer-based data searches of large amounts of texts — on sources they have legal access to. But that provision does not cover researchers at commercial companies, prompting fears that it might hamper public–private research collaborations in the EU. The European Parliament's legal-affairs committee voted in favour of the changes on 20 June; they must now be approved by the Parliament and EU member states.

US ocean policy

US President Donald Trump has written climate change out of the country's plan for managing its coastal waters and the Great Lakes. On 19 June, Trump issued a national ocean policy that emphasizes the development of ocean industries, such as offshore oil and gas drilling. The document replaces a policy that former president Barack Obama released in 2010, following the massive



NSF/PLANET PIX VIA ZUMA

Deepwater Horizon oil spill in the Gulf of Mexico (**pictured**). That plan came under sustained criticism from Republican politicians for its focus on environmental sustainability and stewardship, including several mentions of climate change.

ENERGY

Power initiatives

The Chinese Academy of Sciences (CAS) has launched a large-scale clean-energy project. On 19 June, it announced that 20 of its institutes will join forces to develop dozens of renewable-energy and energy-efficient technologies before 2023. Together, they will aim to help cut pollution from coal-fired power stations by 40–50%, and replace 100 million tonnes of oil and gas usage. China is the world's largest energy consumer; in 2017, it used 608 million tonnes of oil. On 15 June, state-owned nuclear power developer China National Nuclear Corporation announced plans to open a university in the east-coast city of Tianjin to train nuclear engineers who will help to ramp up nuclear production and export nuclear technology. China has an ambitious plan to have 58 gigawatts of nuclear generating capacity in operation by 2020, up from 33.6 gigawatts at the end of 2016.

SPACE

Asteroid plan

The US government has reaffirmed its commitment to finding dangerous near-Earth asteroids and developing ways to deflect them if needed. A 20 June federal report outlines how government agencies will cooperate to upgrade detection capabilities, such as by building new telescopes, and how they can collaborate on new computer models to assess impact hazards. The government has almost tripled its spending on planetary defence in the past few years, from about US$21 million in 2013 to $60 million in 2017. Still, NASA is only about one-third of the way to its goal of tracking the roughly 25,000 near-Earth asteroids that are at least 140 metres across — those big enough to cause serious regional damage if they were to hit the planet.

EVENTS

Scientist sacked

Hungary's reform-minded head of research has been dismissed by the country's prime minister, its innovation ministry announced on 20 June. József Pálinkás, a physicist and former president of the Hungarian Academy of Sciences (HAS), was appointed head of the National Research, Development and Innovation Office in 2015, where he led an expansion in science funding. His dismissal comes one week after a government proposal that would hand the innovation ministry control of much of the HAS's research-funding budget. See page 485 for more.

Marijuana drug

For the first time, the US Food and Drug Administration has approved a treatment containing a component derived from marijuana. In a 25 June announcement, the agency approved Epidiolex, an antiepileptic drug made from cannabidiol. The chemical is found in marijuana but does not cause psychoactive effects. Epidiolex, made by GW Pharmaceuticals in Cambridge, UK, can now be used to treat two epilepsy disorders that manifest in childhood, known as Lennox–Gastaut syndrome and Dravet syndrome. In clinical trials, Epidiolex reduced the number of seizures in some people by more than 50%. The drug is currently under review by the European Medicines Agency.

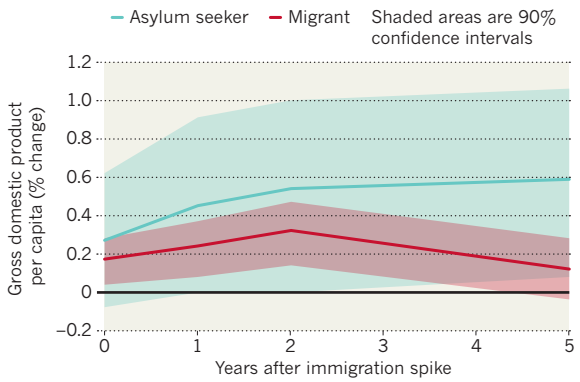
NATURE.COM
For daily news updates see:
www.nature.com/news

TREND WATCH

Asylum seekers and migrants searching for safe havens and opportunities benefit their host nations' economies within 5 years of arrival, suggests an analysis of 30 years of data from 15 countries in Western Europe. Researchers modelled how economic indicators, such as gross domestic product per capita, changed following a spike in immigration. Their model suggested that within two years of an influx of migrants, unemployment rates drop significantly and economic health increases.

THE ECONOMICS OF MIGRATION

A mathematical model based on 30 years of data from 15 Western European countries suggests that migrants and asylum seekers do not burden national economies and instead benefit them.



SOURCE: H. D'ALBIS, E. BOUBTANE & D. COULIBALY *SCI. ADV.* **4**, EAAQ0883 (2018).

NEWS IN FOCUS

RUPAK DE CHOWDHURI/REUTERS

The Indian monsoon can bring damaging floods, but also has a crucial role in ensuring an adequate water supply for people and crops.

ATMOSPHERIC SCIENCE

Mysteries of Indian monsoon probed

Research plane and ships aim to gather the most detailed data yet on rainfall variations.

BY ALEXANDRA WITZE

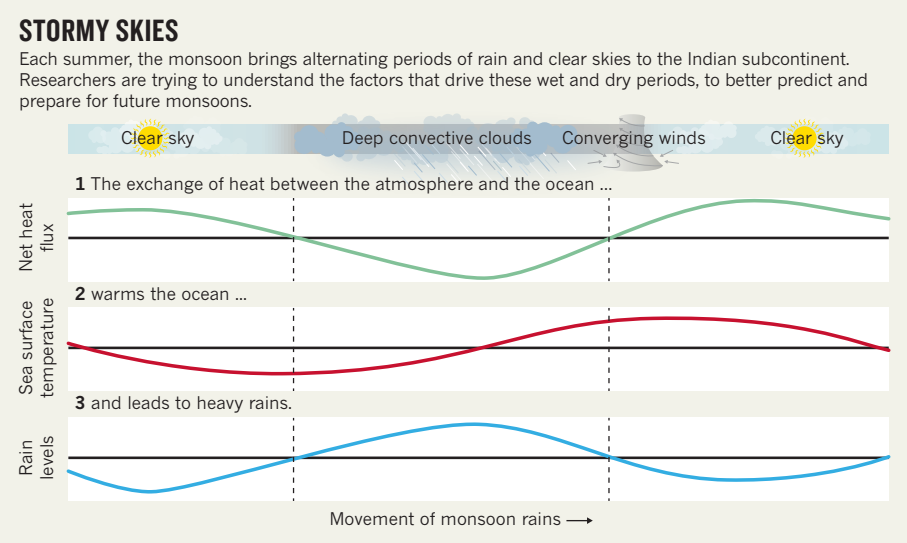
Heavy rains and seven-metre-high waves pummelled the research vessel *Thomas G. Thompson* in the Bay of Bengal this month, routinely drenching the oceanographers on deck. But that was just fine with the scientists. Their entire plan involved getting as wet as possible, in order to directly measure what happens where the air and the sea meet in a summer storm.

The team is part of a multinational group of researchers who are descending on the Indian Ocean this summer to study its seasonal monsoon. They intend to gather the most detailed observations yet on the wet and dry periods that alternate roughly every 10–50 days during the monsoon season, which lasts from June to September.

If modellers could better predict these varying patterns — called monsoon intra-seasonal oscillations, or MISOs — then officials could better prepare for the monsoon each year. That includes timing the planting of crops in concert with the rains, storing water behind dams for hydropower, and preparing for floods and other natural disasters.

"One billion people on the Indian subcontinent depend on planning for water resources," says Harindra Fernando, a mechanical engineer at the University of Notre Dame in Indiana and one of the project's leaders. "When people are waiting for rain, it's important to know when you will get it and when you won't."

MISOs represent the monsoon's 'active' and 'break' periods, in which weeks of heavy rainfall give way to brilliant sunshine before starting all over again. The patterns of rainfall generally track northwards over the Bay ▶



of Bengal, and sometimes veer towards the Indian subcontinent — where they can cause serious damage. In 2017, a powerful MISO brought torrential rain and landslides to Sri Lanka, killing more than 200 people.

Weather and climate models have not been able to accurately predict MISOs[1]. Strong and frequent interactions between the atmosphere and the ocean seem to help get them started[2], as warm ocean waters feed energy into the air above (see 'Stormy skies'). A study published this year suggests that certain ocean processes, such as a type of wave that helps warm the topmost waters, could play a big part in kicking off many MISOs[3]. Having direct measurements from within a MISO will help modellers to pinpoint the exact conditions that drive them, says lead author Jason West, an atmospheric scientist at the University of Colorado Boulder.

The Bay of Bengal project aims to measure the microphysics of energy flows between the ocean and the atmosphere once MISOs are under way. It builds on a long history of international field campaigns to understand the intricacies of the Indian Ocean monsoon, so that neighbouring countries can better prepare for it[4]. Funders of the five-year study, which is known as MISO-BOB, include the Indian Ministry of Earth Sciences and the US Office of Naval Research, working with institutions such as Sri Lanka's National Aquatic Resources Research and Development Agency (NARA).

MISO-BOB's aerial component started on 15 June from an air base in Colombo, Sri Lanka. Scientists have been flying aboard the US Air Force's hurricane-hunter C-130 plane, which carries equipment to measure the properties of clouds and the atmosphere. It is releasing instrument packages called dropsondes that measure temperature, pressure and wind speed as they plummet towards the sea.

The *Thompson*, meanwhile, left Chennai, India, on 4 June for the first of two research legs to gather data on ocean conditions. Scientists on board — mostly students and early-career researchers — have deployed a variety of instruments to measure temperature, salinity, currents and other factors at different depths and locations across the Bay of Bengal. They have also released radiosondes, which are instrument packages carried upward by weather balloons to gather meteorological data.

"We want to observe the conditions across the air–sea interface cleanly, which is a challenging thing to do," says Emily Shroyer, an oceanographer at Oregon State University in Corvallis, who led the first leg. The second leg will take another group of scientists on board and will wrap up by 22 July. This team will travel farther to the south, and an associated group will take measurements near Sri Lanka with the NARA research vessel *Samudrika*, says NARA oceanographer Priyantha Jinadasa.

With enough data and analysis, MISO forecasts could improve in perhaps five or ten years, says team member Debasis Sengupta, a monsoon expert at the Indian Institute of Science in Bangalore.

Project scientists are already planning a second, longer season of field observations for next summer. Details have not yet been finalized, but the team will continue to target how energy flows between the air and the sea during the monsoon. "There's constantly this game going on between the atmosphere and the ocean," says Amit Tandon, an oceanographer at the University of Massachusetts in Dartmouth. "It's every bit as exciting as a World Cup match between two nations." ■

1. Goswami, B. N., Rao, S. A., Sengupta, D. & Chakravorty, S. *Oceanography* **29**, 18–27 (2016).
2. Sharmila, S. *et al. Clim. Dyn.* **41**, 1651–1669 (2013).
3. West, J. B., Han, W. & Li, Y. *J. Geophys. Res. Oceans* https://doi.org/10.1019/2017JC013564 (2018).
4. Mahadevan, A. *et al. Oceanography* **29**, 14–17 (2016).

TRADE

US–China trade war rattles labs

Trump puts tariffs on Chinese technology and China retaliates with taxes on US chemicals.

BY ANDREW SILVER

Scientific research in the United States could become collateral damage in the country's escalating trade dispute with China. Both nations went head-to-head in mid-June over tariffs on a long list of goods that includes lab equipment and reagents. That is likely to increase the cost of scientific research, and the impact could be felt more keenly in US labs.

The latest skirmish in the ongoing trade war between the world's two largest economies began on 15 June, when the United States announced a 25% tax on 818 goods imported from China. The list includes equipment used by scientists, such as basic electrical parts, microscopes and geological-survey devices. President Donald Trump said the tariffs, which will start on 6 July, are intended to reduce China's dominance in industries such as robotics, new materials and information and communications technology, and will level the playing field for US firms. The Trump administration is considering tariffs on a further 284 industrial goods, including chemicals.

A day after the US announcement, China's Ministry of Commerce responded with its own set of tariffs on 545 US products imported to China, which will also start on 6 July. The government will apply taxes in the future to another 114 US imports — including basic chemicals and medical devices, such as magnetic resonance imaging (MRI) machines — although it has not announced a date.

Scientists in the United States were quick to denounce Trump's latest round of tariffs. "I am opposed to these seemingly ad hoc tariffs because it will further stretch the already anaemic scientific research budgets in this country," says Thomas Lapen, a geochemist at the University of Houston, Texas. Equipment

and supplies are the second-largest expense for his research, after paying wages. Lapen says that his costs are likely to increase because the US tariff list includes equipment or parts his team needs, such as electrical motors that drive centrifuges.

Priscilla Cushman, a dark-matter physicist at the University of Minnesota in Minneapolis, says that research deans at US universities should be scrutinizing the list to see whether the taxes will affect their facilities.

SCIENCE SQUEEZE

The tariffs could also cause havoc for large-scale experiments, such as the ADMX dark-matter detector at the University of Washington in Seattle, which is under construction. The project's lead scientist, physicist Leslie Rosenberg, is worried that the equipment his team needs to build experiments — such as tools for power generation and distribution, and machinery that has Chinese electrical components — could be subject to the latest tariffs. "Anyone can see the tariff list, but an official must determine whether any particular procurement falls under the tariff," he says.

Rosenberg thinks that the United States' overall research capability will probably decline under the tariffs.

But other researchers aren't worried. Roberto Refinetti, a biopsychologist who studies biological clocks at Boise State University in Idaho, uses some small Chinese-manufactured equipment for his work, such as infrared motion detectors for monitoring rodents. He doesn't think that the tariffs on Chinese goods will significantly increase the cost of his research, because he purchases that type of equipment infrequently.

The White House and the Office of the United States Trade Representative did not respond to *Nature*'s request for comment on researchers' concerns.

In China, the tariff dispute could increase the cost of standard reagents used in laboratory and medical devices that scientists import from the United States. Ruibang Luo, a bioinformatician at the University of Hong Kong who collaborates with researchers on the mainland, says that if the Chinese government interprets some tariff items literally, the taxes could apply to a broad range of US-made reagents and research devices, including some DNA sequencers.

But Yu Zhou, a researcher at Vassar College in Poughkeepsie, New York, who studies science and technology development in China, says that the tariffs would not have a significant effect on research projects and experiments in China. She says that is because some universities have large enough budgets to absorb increased costs. Researchers could also share more equipment than they do now, or use goods made domestically and from countries other than the United States.

Brian Xu, a toxicologist for the scientific consulting firm ACTA in Washington DC, which works with businesses in China, agrees that China's proposal to place tariffs on US chemicals and scientific equipment is unlikely to have a major effect on Chinese research. He notes that scientists there import only a small amount of US-made chemicals, and that infrequently replaced scientific equipment from other countries, such as Japan and Germany, is of comparable quality and cost.

But the latest round of tariffs might not be the last. On 18 June, Trump threatened to impose additional tariffs on Chinese goods if the country does not rescind its tariffs and create a more balanced trade relationship with the United States. ■

SPACE

Japanese mission reaches unexplored asteroid Ryugu

Hayabusa-2 will release four landing probes before touching down to collect samples.

BY DAVIDE CASTELVECCHI

After travelling for three and a half years, the Japanese spacecraft Hayabusa-2 this week makes its final approach to the asteroid Ryugu. The probe will release landers on the space rock's surface later this year and bring a precious sample back to terrestrial labs in 2020. It is already giving planetary scientists their closest-ever view of a mysterious kind of asteroid.

The Japan Aerospace Exploration Agency (JAXA) last week released grainy pictures from a distance of around 300 kilometres away, revealing that Ryugu — an asteroid of a common but little-studied type — looks similar to a spinning top.

This week, a much more detailed picture, from 40 kilometres away, showed a surface strewn with large boulders. Hayabusa-2 will continue to inch towards the asteroid until it is about 10 kilometres away, which JAXA expects will happen around 27 June. Ryugu's orbit cuts between those of Earth and Mars.

"From a distance, Ryugu initially appeared round, then gradually turned into a square before becoming a beautiful shape similar to fluorite, known as the 'firefly stone' in Japanese," project manager Yuichi Tsuda said in a 25 June statement.

Launched in December 2014, the probe is a follow-up to — and near-clone of — Hayabusa, which explored the asteroid Itokawa starting in 2005. Hayabusa was the first mission to return an asteroid sample to Earth. Ryugu is about 1 kilometre across — around 3 times wider than Itokawa but one-quarter the size of the comet 67P/Churyumov–Gerasimenko, which the European Space Agency's Rosetta probe visited between 2014 and 2016.

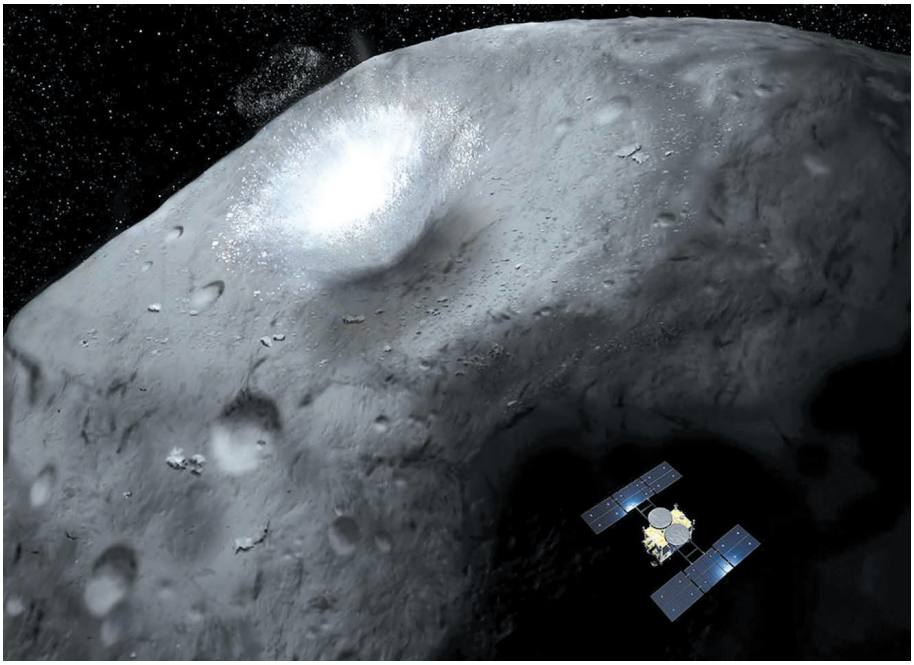
"Ryugu initially appeared round, before becoming a beautiful shape similar to fluorite."

Ryugu is a 'C-type' asteroid, which has a darker surface than does Itokawa. In 1997, a NASA mission called NEAR Shoemaker made a fly-by of a C-type asteroid from a distance of more than 1,000 kilometres. Hayabusa-2 is the only spacecraft to have come this close to a C-type asteroid, says Lucy McFadden, a planetary scientist at NASA's Goddard Space Flight Center in Greenbelt, Maryland.

"We don't know much about C-type asteroids," says McFadden. But they are expected to have a composition similar to that of the early Solar System. In particular, Hayabusa-2 will determine whether the darkness of Ryugu's surface is due to it being rich in carbon — as is often assumed — or to small, metallic particles such as magnetite.

Chemical and isotopic analyses of the rock — to be done in space by Hayabusa-2's landers and then in terrestrial labs — could help to explain the origins of Earth and, particularly, its water. Many researchers think that Earth's oceans formed from a bombardment of water-rich asteroids or comets.

Among the first measurements Hayabusa-2 made was one of Ryugu's rotational period, or time it takes to make one turn on its own ▶



JAXA

An artist's impression of Hayabusa-2.

axis, which is about 7.5 hours. This is good news, because a much faster rotation could have made it harder to approach the surface, says mission manager Makoto Yoshikawa of JAXA's Institute of Space and Astronautical Science in Sagamihara. But its shape was surprising, he says, because it has a bulge around the equator, something that is usually associated with much faster-spinning objects, Yoshikawa adds.

Hayabusa-2's most important task right now is to pinpoint its own position using laser ranging so that it can manoeuvre accordingly. "We want to know the exact distance of the spacecraft to the asteroid," Yoshikawa says. Also crucial is to map the asteroid surface using its on-board camera and infrared spectrometer. Temperature variations will hint at the composition of the surface. All of these data will be crucial for deciding where to release MASCOT, the shoebox-sized lander that will probe the asteroid, and the three other small probes carried by the mothership.

"We will use the information we get from the mother spacecraft to do landing-site selection," says MASCOT payload manager Stephan Ulamec of the German Aerospace Center in Cologne. Ulamec was also project manager for Philae, a probe that Rosetta released onto the surface of 67P/Churyumov–Gerasimenko. That approach took several hours because Rosetta was orbiting the comet and Philae had to spiral down to its surface.

Hayabusa-2 will simply hover over Ryugu — using its own gentle ion engines to counteract the asteroid's gravitational attraction — and release MASCOT straight down. Some time in October, the lander will make a soft touchdown. After MASCOT settles on the surface, an internal mechanism will straighten the lander up so it can use its on-board instruments and communicate with Hayabusa-2.

MASCOT carries no solar panels and its batteries are expected to last only a few hours. The team will meet in Toulouse, France, in mid-August to make the final selection for the landing site of MASCOT and its companions.

Meanwhile, Hayabusa-2 will make its own, brief soft landings to collect samples of the asteroid's surface. Then, in late 2019, it will head back to Earth, a journey expected to last a year. Compared to the more daring manoeuvres to reach the asteroid's surface, the current part of the mission is relatively low risk, says Yoshikawa. But as the craft approaches Ryugu, his team has already kicked into high gear, he adds: "I do not have much time for sleep." ■

CLIMATE CHANGE

Methane leaks from US gas fields dwarf official estimates

Latest study suggests that emissions could be coming from faulty equipment.

BY GIORGIA GUGLIELMI

Methane leaks from the US oil and gas industry are 60% greater than official estimates, according to an analysis of previously reported data and new airborne measurements.

Because methane is a potent greenhouse gas, scientists say that the unaccounted-for emissions could have significant impacts on the climate.

The analysis[1], published on 21 June in *Science*, is one of the most comprehensive looks yet at methane output from US oil and gas production, and reinforces previous studies that suggested emissions outpaced government estimates. That research prompted the government to develop regulations that would restrict methane emissions from oil and gas production — rules that US President Donald Trump is now attempting to roll back.

The latest study shows that the US oil and gas supply chain emits about 13 million tonnes of methane, the main component of natural gas, every year. That's much higher than the US Environmental Protection Agency (EPA) estimate of about 8 million tonnes.

This discrepancy probably stems from the fact that the EPA's emissions surveys miss potential sources of methane leaks, such as faulty equipment at oil and gas facilities, says study leader Ramón Alvarez, an atmospheric chemist at the Environmental Defense Fund, a non-profit group in Austin, Texas.

Methane warms the planet 80 times as much as carbon dioxide does over the first 20 years after it is released. And atmospheric methane contributes to about 25% of global warming, Alvarez says. "That's a significant amount."

If left unchecked, he says, methane emissions from the oil and gas industry could erode the potential climate benefits of using natural gas, which releases much less CO_2 and other toxic pollutants than coal does when it is burned.

The latest study comes one year after the EPA announced a delay of the rule that would restrict methane emissions produced by oil and gas drilling operations. The policy,

introduced under former president Barack Obama, will not take effect until 2019.

Before 2012, published estimates of the US methane leakage rate ranged from 1% to about 8%, Alvarez says, and the lack of consensus pushed scientists to improve measurements of those rates in subsequent years. Alvarez and his team pooled data from some of these studies — many of which quantified emissions at individual facilities — and validated the measurements using aircraft surveys. The team covered regions accounting for about 30% of US gas production.

The researchers then extrapolated the figures to estimate methane leaks at the national level. The team found a leakage rate of 2.3% in 2015, compared with the 1.4% estimate from the EPA. The gas is escaping through holes in the production system, and it adds up to a lot of emissions, says Alvarez.

The findings reduce the uncertainty around the magnitude of US methane emissions, says Daniel Zimmerle, an energy researcher at Colorado State University in Fort Collins. "But I would be surprised if this would be the final word on the topic."

Because of methane's warming potential, a leak rate of 2.3% is concerning, says Robert Howarth, an Earth-systems scientist at Cornell University in Ithaca, New York. But he cautions that the study might have underestimated the actual leak rate of methane. Howarth notes that the measurements scientists used include some obtained with an instrument that — according to the device's inventor — produces systematically low numbers[2].

What's more, Howarth says, the researchers didn't look at the emissions from systems that distribute gas to urban areas, which studies suggest are considerable[3].

But Alvarez looks on the bright side. Because a substantial proportion of these leaks is probably due to faulty equipment, he sees a "tremendous opportunity" to reduce methane emissions by developing systems to quickly detect malfunctions at oil and gas facilities, and by identifying overlooked ways in which the greenhouse gas escapes into the atmosphere. ■

1. Alvarez, R. *et al. Science* https://doi.org/10.1126/science.aar7204 (2018).
2. Howard, T. *Environ. Sci. Technol.* **49**, 3981–3982 (2015).
3. McKain, K. *et al. Proc. Natl Acad. Sci. USA* **112**, 1941–1946 (2015).

MICROBIOLOGY

Bacteria deliver gene therapies

Engineered strains of Escherichia coli *and other microbes are being tested in people as treatments for a slew of illnesses.*

BY SARA REARDON

People often take medicines to rid themselves of problem bacteria. Now, a counter-intuitive approach — turning genetically modified bacteria into medicines — is gaining ground.

Several companies are testing whether engineered bacteria can treat conditions that affect the brain, liver and other organs — and even kill other, harmful microbes. But although US regulators have approved trials of several types of engineered bacterium as a form of gene therapy, questions remain about whether microbes' ability to share DNA with one another will create long-term safety risks.

The idea of using bacteria to deliver gene therapies first surfaced in the 1990s, but early clinical trials met with mixed results. Interest in the approach has increased in recent years amid mounting evidence that the bacteria that live in the body — the microbiome — can influence human health. Researchers are looking to treat disease by modifying microbes that are normally found in people or the foods they consume.

"The microbes are extremely smart and they know how to survive."

Matthew Chang, a synthetic biologist at the National University of Singapore, says that genetically modified bacteria have the potential to treat many diseases. His group is engineering the gut bacteria *Escherichia coli* and *Lactobacillus* to recognize and destroy harmful microbes[1]. "It's a rapidly growing area," says Chang, who adds that he is in talks with regulators in Singapore about starting clinical trials.

One strain of research is aimed at treating the genetic disorder phenylketonuria. People with the condition are deficient in an enzyme that breaks down the amino acid phenylalanine, which causes neurological damage if it builds up in the body. At the American Society for Microbiology's annual meeting in Atlanta, Georgia, earlier this month, researchers from the biotechnology firm Synlogic in Cambridge, Massachusetts, reported that *E. coli* modified to produce an enzyme that degrades phenylalanine, and a protein that moves it from blood to cells, reduced levels of the amino acid in monkeys' blood by more than half compared with animals in a control group. The company started clinical trials in healthy human volunteers in April, and will begin testing the bacteria in people with phenylketonuria as soon as it concludes that the therapy is safe.

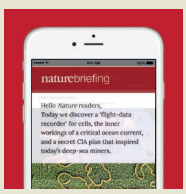
Another firm, Intrexon of Germantown, Maryland, has altered *Lactococcus lactis*, a bacterium used in cheese production, to make a protein that protects skin's outer layers. One clinical trial that has enrolled about 200 people with cancer is testing whether an *L. lactis* mouthwash can prevent oral sores that are a side effect of chemotherapy. In July, the company will begin dosing people who have diabetes with a different form of *L. lactis* that produces both the precursor to human insulin and an immune protein that enhances cells' ability to respond to insulin.

Both Intrexon and Synlogic have engineered their bacteria to reduce their likelihood of establishing colonies in the body — which means that patients would have to take the microbes regularly. But other companies are pursuing treatments that would create colonies of transgenic bacteria in the body.

The biotechnology firm Osel in Mountain View, California, plans to seek US ▶

MORE ONLINE

NATURE BRIEFING



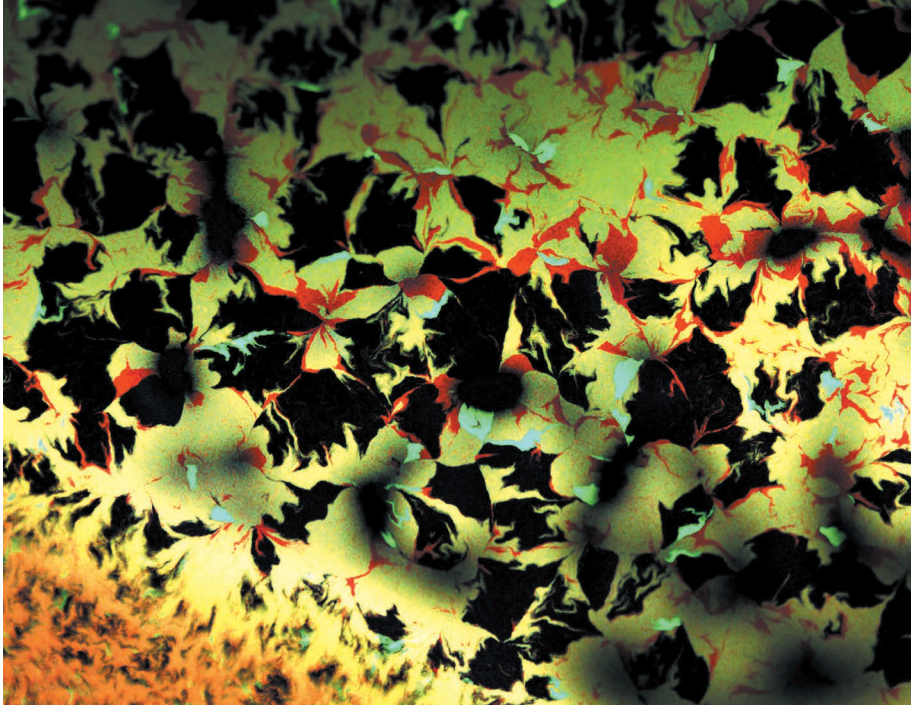
Save time — get the daily *Nature* Briefing direct to your inbox go.nature.com/savetime

MORE NEWS

- African scientists launch their own preprint sever go.nature.com/2mqzso6
- Ancient gibbon from Chinese tomb may be first ape to go extinct since the Ice Age go.nature.com/2kmvqeo
- Silica nanograins form perfect 12-sided cages go.nature.com/2mmdxtk

NATURE PODCAST

Infant air pollution deaths; stressed brains; and diagnosing sick plants from afar nature.com/nature/podcast



The *Escherichia coli* bacteria is being developed as a vehicle for gene therapy in people.

FERNAN FEDERICI AND JIM HASELOFF/WELLCOME COLL./CC BY

government approval this year for a *Lactobacillus* that has been engineered to prevent HIV transmission. Studies have shown that naturally high levels of *Lactobacillus* in the vagina can help to protect women against HIV[2]. Osel is attempting to enhance the bacterium's protective properties by modifying it to make a human protein that prevents HIV infection.

Challenges remain before these engineered bacteria can enter the market. There is a risk that the microbes could pass the human genes they carry to other bacteria in the body, with unknown consequences. Several companies have attempted to prevent this exchange by altering the chromosomes of a bacterium, rather than its plasmids — tiny pieces of DNA that bacteria pass back and forth. They have also built in biological 'kill switches' to prevent the microbes from surviving outside the body.

This strategy can fail, however. A group led by immunologist Simon Carding of the University of East Anglia in Norwich, UK, engineered[3] *Bacteroides ovatus* to treat colitis, an inflammation of the intestine. The group also edited the bacterium's chromosome to make it dependent on a molecule produced by naturally occurring gut bacteria.

But just 72 hours after the scientists fed the bacteria to mice, they found that *B. ovatus* had passed its modified gene to other microbes in the animals' guts — and acquired genes that allowed it to live without the molecule.

The experience caused Carding to abandon efforts to develop bacteria as therapies. "It's potentially harmful if it's not properly controlled," he says.

Synlogic, Osel and other companies say they have never observed this type of gene transfer, but agree that it is possible. "The microbes are extremely smart and they know how to survive," says Chang. It remains to be seen, he adds, whether engineering bacteria to colonize the body or die out quickly is a better approach — but the answer could emerge as the current set of clinical trials wraps up in the next few years. ■

1. Hwang, I. Y. *et al. Nature Commun.* **8**, 15028 (2017).
2. Gosmann, C. *et al. Immunity* **46**, 29–37 (2017).
3. Wegmann, U., Carvalho, A. L., Stocks, M. & Carding, S. R. *Sci. Rep.* **7**, 2294 (2017).

SPECIAL ISSUE

Hubs of East Asian science

A special issue explores the booming research landscapes of Hong Kong, Malaysia, Singapore, South Korea and Taiwan.

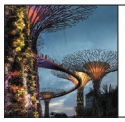
Science is booming in many parts of Asia — but it is too easy to focus only on the giant economies and overlook other research powers. This special issue shines the spotlight on five strong science centres in East Asia: Hong Kong, Malaysia, Singapore, South Korea and Taiwan. This group is incredibly diverse in priorities and approaches, but all see science and technology as keys to their future. Together, they have an expanding role in the global research enterprise.

An infographic on page 500 explores the investments that members of this group have made in research and development, and how they have raised their international standing in spending and research output. The breadth of their science and technology programmes is illustrated by profiles of ten remarkable researchers who are advancing science — ranging from genome editing to green-energy development — and supporting their communities (see page 502).

Two Comment articles chart paths for research in South Korea and Malaysia. Physicist and presidential science adviser Han Woong Yeom argues on page 511 that South Korea must spend more on basic research and infrastructure while also addressing public concerns such as health care and air pollution. Asma Ismail, president of the Academy of Sciences Malaysia, issues a call on page 514 for scientists to further progress in social as well as economic areas, from green technology to the Halal economy. And an Editorial on page 485 urges science officials across the world to address local needs.

The growth of science in this region is a lure for scientists. A Careers Feature on page 625 traces the path of four researchers who left East Asia to pursue PhDs and postdoctoral research, but have now returned to develop their research.

With their substantial investments and strengths, these five research centres in East Asia are working to secure their future as major forces in the global science landscape. ■



HUBS OF ASIAN SCIENCE
A *Nature* collection
nature.com/collections/asianhubs

FIVE IN ASIA

HONG KONG, MALAYSIA, SINGAPORE, SOUTH KOREA AND TAIWAN ARE INVESTING HEAVILY IN RESEARCH AS AN ENGINE FOR GROWTH.

BY RICHARD VAN NOORDEN

Which economies invest the most in research and development (R&D)? The answer might not be what you think. South Korea ploughs a whopping 4.24% of its gross domestic product (GDP) into science and technology — neck and neck with Israel, and putting much of Europe and the United States to shame. Taiwan also invests heavily, beating science heavyweight Japan in 2016 in terms of the share of its economy devoted to R&D.

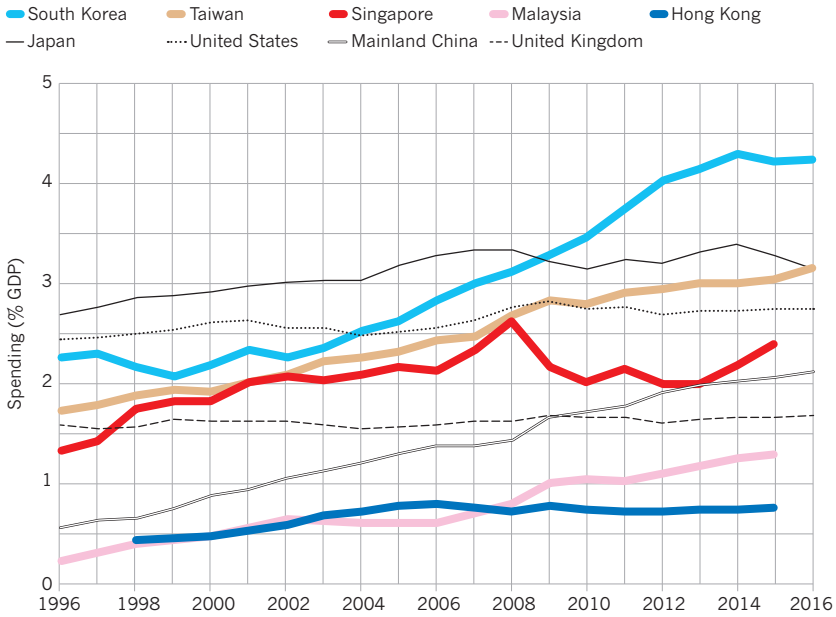
In East Asia, several science powerhouses are investing strongly in science. Although mainland China and Japan get much of the attention here — they are the area's biggest economies and have giant scientific workforces — South Korea, Taiwan, Singapore and Hong Kong have established themselves as strong supporters of research, and Malaysia is fast growing its scientific output.

HUBS OF ASIAN SCIENCE
A *Nature* collection
nature.com/collections/asianhubs

SPENDING

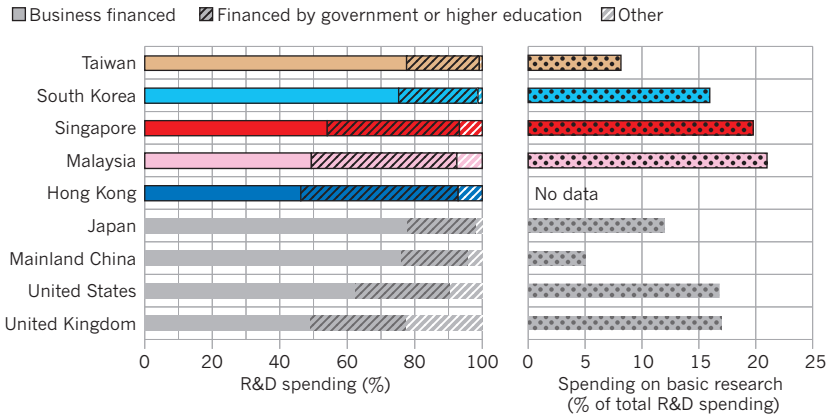
Research and development (R&D) investment is rising rapidly in South Korea, Taiwan and Malaysia — albeit from different bases. In two decades, South Korea has close to doubled the share of its economy spent on research. Taiwan's proportion is not far behind, and it overtook Japan in 2016. Singapore's spending was keeping pace with Taiwan's, but has dropped off because of a decline in business R&D spending. Only Hong Kong's investment has plateaued in the past decade or so.

SOURCE: UNESCO/OECD



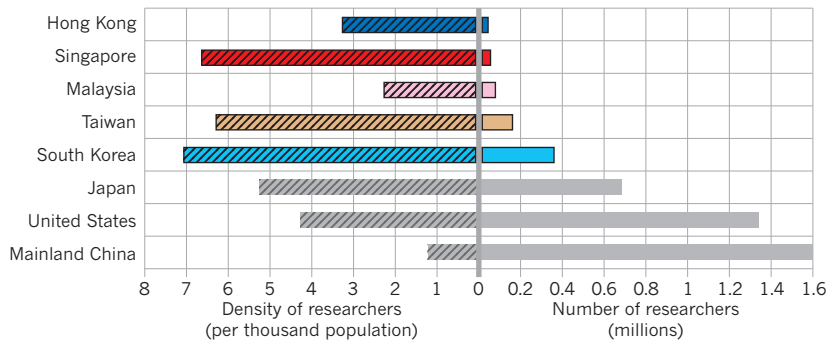
BUSINESS VERSUS GOVERNMENT

Taiwan and South Korea's R&D investment comes mainly from the business sector, and the proportions invested in basic research are accordingly lower. In Singapore, Malaysia and Hong Kong, businesses provide around half the R&D financing, more like the United States and United Kingdom.



SCIENCE WORKFORCE

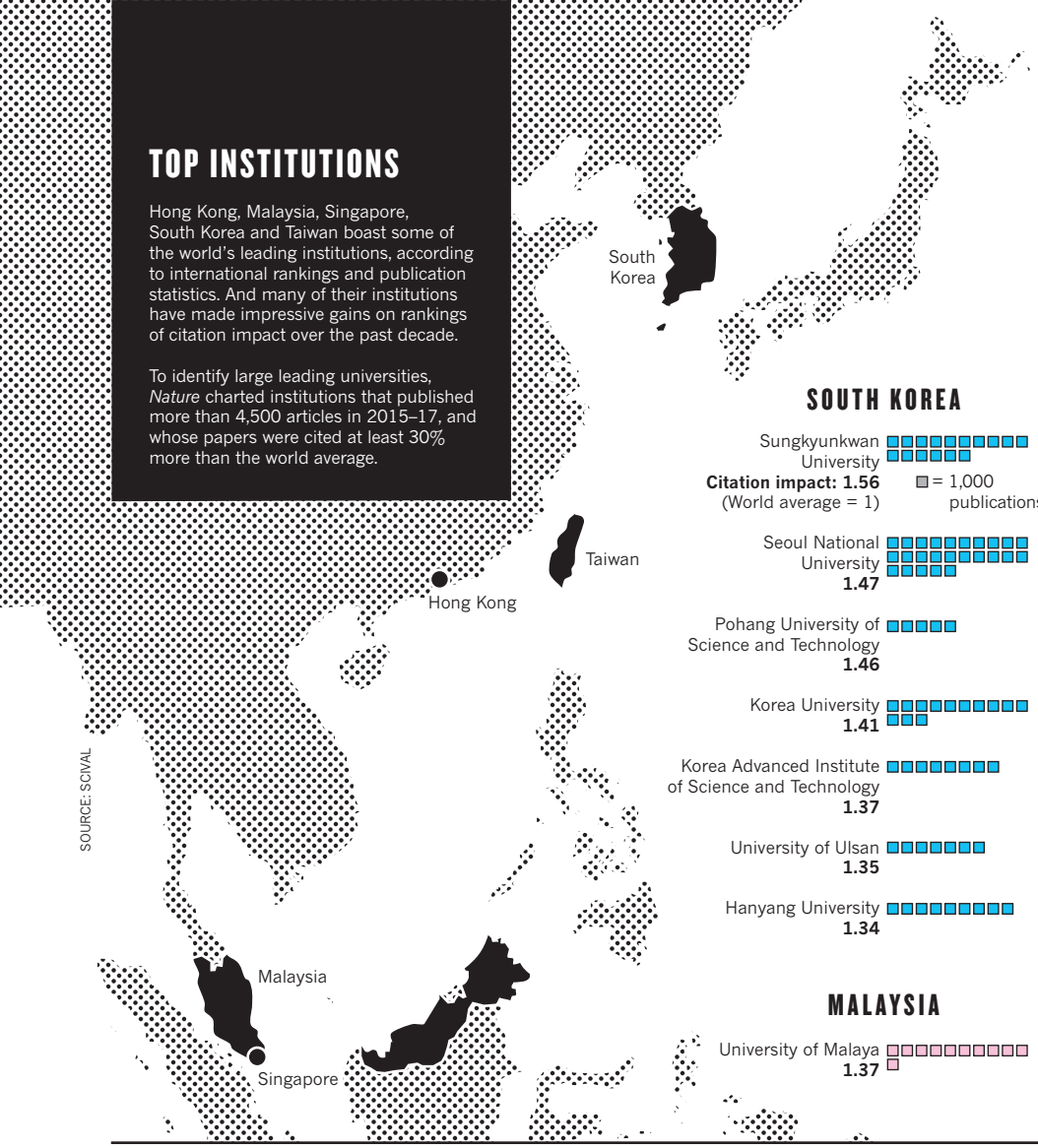
After mainland China and Japan, South Korea has the most researchers in East Asia — and Malaysia has grown its science workforce to a point where it is now ahead of Singapore and Hong Kong. But in terms of researcher density, South Korea, Taiwan and Singapore stand out.



TOP INSTITUTIONS

Hong Kong, Malaysia, Singapore, South Korea and Taiwan boast some of the world's leading institutions, according to international rankings and publication statistics. And many of their institutions have made impressive gains on rankings of citation impact over the past decade.

To identify large leading universities, *Nature* charted institutions that published more than 4,500 articles in 2015–17, and whose papers were cited at least 30% more than the world average.



SOURCE: SCIVAL

SOUTH KOREA

Sungkyunkwan University
Citation impact: 1.56
(World average = 1)
■ = 1,000 publications

Seoul National University
1.47

Pohang University of Science and Technology
1.46

Korea University
1.41

Korea Advanced Institute of Science and Technology
1.37

University of Ulsan
1.35

Hanyang University
1.34

MALAYSIA

University of Malaya
1.37

HONG KONG

Hong Kong University of Science and Technology
2.00

Chinese University of Hong Kong
1.94

City University of Hong Kong
1.90

The University of Hong Kong
1.84

Hong Kong Polytechnic University
1.73

TAIWAN

National Tsing Hua University
1.65

National Taiwan University
1.45

Academia Sinica
1.37

SINGAPORE

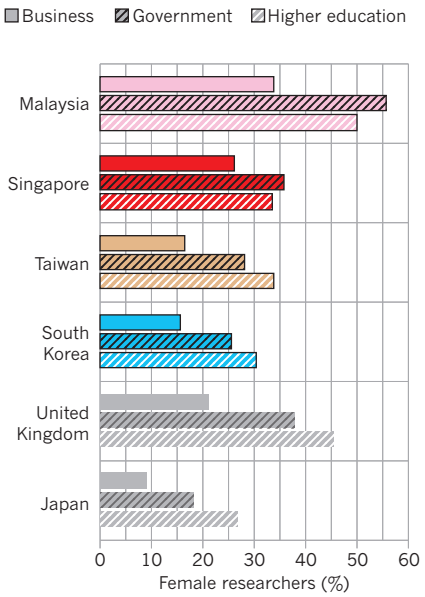
Agency for Science, Technology and Research
1.96

Nanyang Technological University
1.84

National University of Singapore
1.75

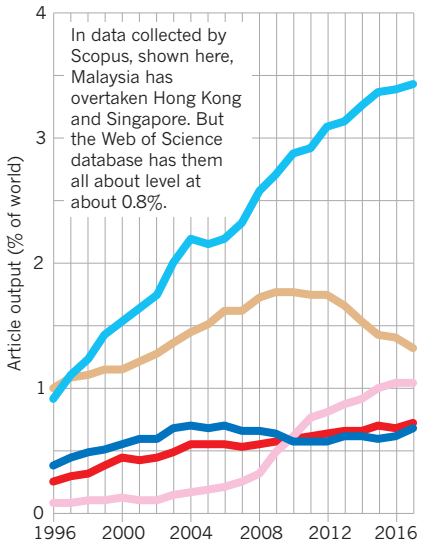
FEMALE RESEARCHERS

Female researchers are under-represented in many Asian economies, as they are across the world. But almost half of Malaysia's researchers are female, and the United Nations says that Malaysia is a world leader in encouraging girls and women to participate in science.



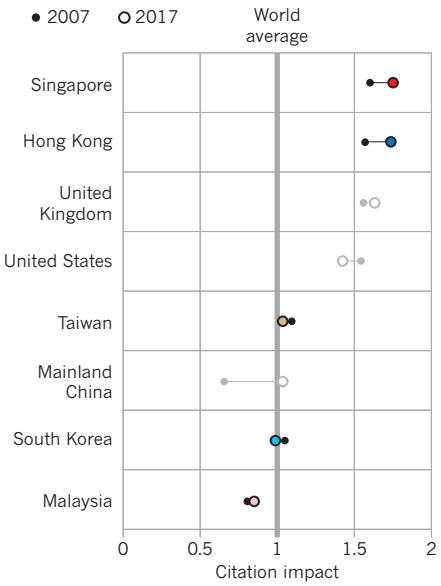
ARTICLE OUTPUT

South Korea's research output had soared to some 65,000 research articles in the Scopus database last year. (By comparison, mainland China produced more than 414,000 articles and scientists in Japan published 89,000). Taiwan's output is dipping as a proportion of the world's research, but Malaysia's volume is rising fast.



CITATION IMPACT

Singapore and Hong Kong have stretched their lead over the United States and United Kingdom in terms of the average scholarly impact of their publications — with normalized citations far above the world mean. One reason is that these economies have very high rates of international collaboration, which is linked to increased citations.



SOURCES: FEMALE RESEARCHERS: UNESCO/OECD; ARTICLE OUTPUT, CITATION IMPACT: SCIVAL

Science stars of EAST ASIA

From artificial intelligence to infectious diseases, top researchers in Hong Kong, Malaysia, Singapore, South Korea and Taiwan are making big impacts on the global stage.

HUBS OF ASIAN SCIENCE
A *Nature* collection
nature.com/collections/asianhubs

AMRITA CHANDRADAS FOR *NATURE*

CANCER DETECTIVE

A geneticist works to improve breast-cancer screening among women in Singapore.

BY SANDY ONG

Jingmei Li does some of her best thinking underwater. "Research is a lot like diving because it's about looking for things that are hiding in plain sight," says Li, an avid diver and a human geneticist at the Genome Institute of Singapore. In her research, Li sifts through information collected from women with breast cancer to detect the risk factors that make them more susceptible to developing the disease. "All the data are there — it's about forming a research question," she says.

To make predictions about whether an individual has a high risk for breast cancer, Li looks for genetic markers as well as lifestyle factors, such as obesity, that are linked to an increased risk of cancer. She also incorporates a third component into her risk assessment, using mammogram images to study breast density. Searching through this combination of data is unusual, says medical epidemiologist Per Hall at Stockholm's Karolinska Institute, who was Li's PhD supervisor. "She's one of the new generation of epidemiologists," says Hall, who adds that risk models are getting better but still require a lot more work.

Li studied the genetic profiles of hundreds of thousands of European women to search for disease markers. Specifically, she looked at the most common type of genetic variation among people — changes in the individual 'letters' in DNA, known as single nucleotide polymorphisms (SNPs). Unlike mutations in the *BRCA1* and *BRCA2* genes, which can substantially increase the risk of developing breast cancer, variations in individual SNPs do not raise a woman's chance of getting the disease by a large margin. However, collective changes to numerous SNPs can elevate a woman's risk significantly (K. Michailidou *et al. Nature* **551**, 92–94; 2017).

After moving back to Singapore in 2017, Li switched her focus to working with local data. Breast cancer is the most common cancer affecting Singaporean women, but they are reluctant to discuss the disease and to get regular screenings, says Li. "I think genetics is the way to go," she says. "It doesn't solve everything, but at least you can identify the high-risk population." Physicians can then concentrate on getting them screened for the disease.

"Personally, I have always thought that if you really want to implement a change, Singapore is the place to do it because the country is so small and compact," Li says. She is also keen to tackle breast infections in older women, a problem more commonly observed in Asia than in the West. "The whole purpose of my work," she says, "is to improve women's lives." ■

VERONICA SANCHIS FOR *NATURE*

THE SENTINEL

An infectious-disease specialist in Hong Kong battles emerging pathogens.

BY DAVID CYRANOSKI

In late 2002, a mysterious and debilitating respiratory illness emerged in China's Guangdong province, and then spread to Hong Kong and eventually the world. Severe acute respiratory syndrome, or SARS, as the disease came to be called, "came out of nowhere", says Malik Peiris. It eventually killed almost 800 people in more than two dozen countries.

SARS' toll could have been much worse had it not been for Peiris, one of the top specialists in emerging diseases. Despite concerns that the virus could kill them or their families, Peiris and his team quickly pinned down the cause and developed measures to control it.

Peiris, a virologist at the University of Hong Kong's School of Public Health, is a much-sought collaborator who can navigate political issues that arise during outbreaks. "He is quiet, and his personality is very conducive to collaboration between international scientists of many cultures," says Robert Webster, a virologist at St Jude's Children's Research Hospital in Memphis, Tennessee, and a pioneering researcher of avian influenza.

Before coming to Hong Kong, Peiris bounced back and forth between Sri Lanka, where he was born and completed an undergraduate degree, and the United Kingdom, where he did postdoctoral research and a stint with the National Health Service.

In 1995, Peiris took on the challenge of establishing a clinical-virology

AIR WARDEN

A Taiwanese pollution expert tackles indoor threats.

BY NICKY PHILLIPS

In 1999, Huey-Jen Jenny Su approached Taiwanese politicians with a radical proposal. Su, a specialist in indoor air pollution, had led a team that had spent more than a decade measuring levels of contaminants in people's homes and workplaces and then tracking the health impacts. Then she brought the evidence to the government and asked it to consider setting air-quality standards for indoor environments. In 2005, Taiwan proposed limits for the levels of certain indoor pollutants, and in 2012 the government introduced laws to regulate indoor air quality.

"I continue to be proud that when the standard was issued we had a huge mass of scientific evidence that was so solid that no one could challenge the decision," says Su, who in 2015 became the first woman president of the National Cheng Kung University in Tainan.

Su has a talent for getting people to stop posturing and start focusing on the task at hand, says John Spengler, an environmental-health researcher at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts, who was Su's doctoral supervisor. "That's why she's been so effective."

Su has long been interested in how pollution affects people. After completing her PhD, she spent a year working on a major US government research project that sent her into many homes, where she saw first-hand the poor living conditions that some people endure. This field experience, she says, "gave me a lot of inspiration after I returned to Taiwan about how to effectively link scientific findings to policy".

Back home, Su and her team revealed how Taiwan's climate generated a particular cocktail of indoor pollutants. They found that many homes in the hot and humid south had high concentrations of allergens, such as dust mites and airborne fungi, above the levels known to cause respiratory problems.

They also found that Taiwan's subtropical environment had higher concentrations of endotoxin — a component of bacterial cell walls linked to asthma severity — than did more temperate climates. Su and her colleagues later demonstrated that living in a home with visible mould and water damage was associated with asthma in adults and children (N.-Y. Hsu *et al. Arch. Environ. Occup. Health* **67**, 155–162; 2012).

Since becoming university president, Su has focused her research on how data on air pollution and its health impacts can help efforts to make society more resilient to global warming. "What choices do we have to help people and our built environment be more adaptive to the changing climate?" she says.

Su is proud that the government standards she helped to develop are protecting people's health, especially those with limited financial resources. "Someone who is committed to public health should always keep in mind those people who are disadvantaged," she says. "That has been my core value and principle all these years." ■

laboratory in Hong Kong. He was put to the test two years later, when Hong Kong recorded the first known case of the H5 strain of bird flu jumping to people. Eventually, 18 became infected and 6 died. The outbreak terrified infectious-disease experts because H5N1 was known for wreaking havoc on bird populations.

Peiris made his name with H5N1. He discovered how the virus kills people by triggering an overreaction in the immune system — known as a cytokine storm — that attacks the body's organs (C. Y. Cheung *et al. Lancet* **360**, 1831–1837; 2002) . He also showed that interventions in live-poultry markets — such as closing the market for one or two days every month, or removing the poultry overnight — dramatically reduced the spread of the virus in the market, thereby reducing risk to people.

The H5N1 scare helped Peiris and others to prepare Hong Kong for future outbreaks. "The system set up in response to bird flu is what helped us to deal with SARS," he says.

Six years later, Peiris' scientific reputation burgeoned through the SARS crisis. He led a team that discovered that SARS was caused by a coronavirus, and developed strategies to contain the virus.

Lately, Peiris has shifted his attention to Middle East respiratory syndrome (MERS), another coronavirus disease with pandemic potential, which he says isn't getting enough attention. Evidence suggests that the virus has been crossing from camels to humans — mainly in Saudi Arabia — in a way that reminds Peiris of local SARS outbreaks before the disease spread globally. "It has the greatest capacity for huge global impact," he says.

Although his scientific reach is now global, Peiris says his location in Hong Kong remains crucial to catching emerging threats. The city has long been a crossroads of diseases that go global. "For a career in the type of things I'm doing," he says, "it's a perfect place to be." ■

AHMAD YUSNI FOR *NATURE*

POWER PLAYER

A chemical engineer turns waste into fuel.

BY YAO-HUA LAW

Suzana Yusup didn't have a choice in her course of study when she arrived at the University of Leeds, UK, in 1992. She had received a scholarship from the Malaysian government to do an undergraduate degree in chemical engineering, but, she says, "I didn't know what chemical engineering was".

She quickly grew fond of the field for its mix of science and practical applications — so fond that she received a PhD in chemical engineering from the University of Bradford, UK, in 1998. The eldest of seven children, Yusup is the only academic in her family. "My parents wanted me to be a doctor, any doctor, but they knew I was afraid of blood. So for them, this [degree] is something to be very proud of."

JOE RUSSO FOR *NATURE*

Yusup now heads the Center for Biofuel and Biochemical Research at Universiti Teknologi Petronas (UTP) in Perak, Malaysia. She joined the university in 2001 and began exploring Malaysia's rich plant resources to produce biofuels. Much of her work has focused on creating fuels from biomass waste, such as used cooking oil, rubber-seed oil and discarded distillate from palm-oil refineries.

She has long explored ways in which green technology can help the environment and society. When she learnt that rice farmers around her university were struggling with pests, weeds and the health threat of chemical pesticides, Yusup started developing safer biopesticides based on compounds from plants. She adapted the techniques and facilities in her biofuel lab to produce the pesticides. Yusup has won many honours, including Malaysia's Rising Star Award in the category of Highly Cited Review in 2016 and the Women in Science Award in 2017.

Yusup's latest research interests reflect her hobby of gardening. In particular, she worries that she might be too weak to plough the soil after she retires. "I am an engineer and I should do something about it," says Yusup. She developed a hydroponic system that is soil-free and requires no ploughing, then got her engineering students to teach it to schoolchildren. Her students complained, but quickly turned enthusiastic after their first harvest. When researchers see the impact they have on others, it helps motivate them to find solutions to problems, says Yusup. "That's what keeps me going." ■

GRAPHENE GROWER

A physical chemist takes a trick from frogs to develop better ways to grow graphene.

BY SANDY ONG

As a boy growing up in Singapore, Loh Kian Ping fell in love with science after seeing specimens of dissected frogs. Decades later at the National University of Singapore, the physical chemist once again drew inspiration from frogs — this time, to make one of his key discoveries. By mimicking how the amphibians latch on to leaves, Loh developed a method to connect graphene to silicon wafers, which could help in industrial applications such as improving optical communications.

With its single layer of carbon atoms, graphene is hailed by the microelectronics industry for its thinness and its ability to conduct electricity. But growing graphene on a surface is tricky. If the conditions are wrong, the thin layer of carbon peels off and floats away in solution. Or too many carbon layers grow and create graphite instead.

Loh and his colleagues devised a way to grow graphene on copper-coated silicon wafers, then etch away the copper (L. Gao *et al. Nature* **505**, 190–194; 2014). To prevent graphene from washing away as well, Loh made special pretreated silicon wafers that attach to the newly grown graphene through capillary bridges — the same microstructures that allow frogs to adhere to lotus leaves. "It's a biomimetic way of attaching graphene to silicon wafers," says Loh.

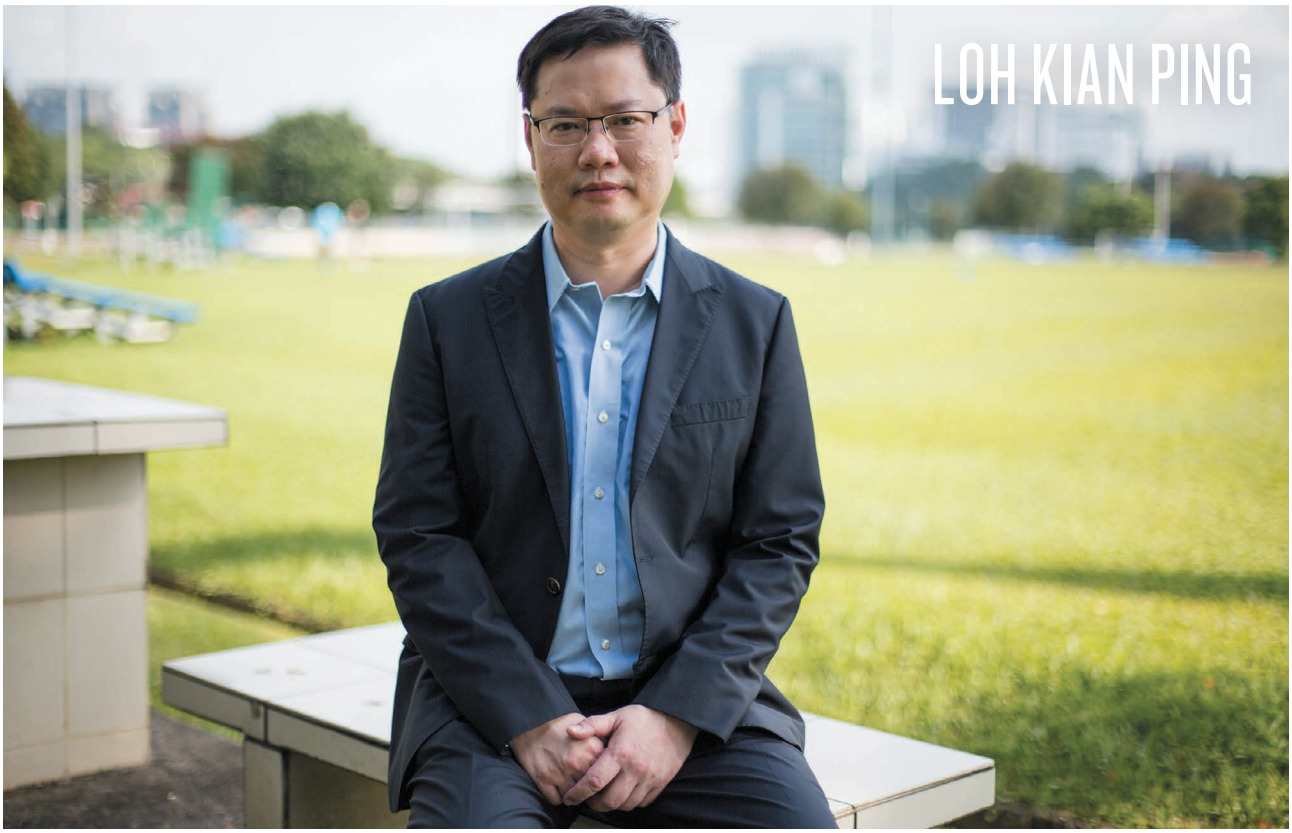
Successfully growing graphene on silicon microchips is a crucial step towards using the material on a commercial scale in industry to improve the performance and cost of electronic devices, says chemist Xiangfeng Duan at the University of California, Los Angeles. Loh's achievement, he says, helped to address a key bottleneck.

Loh has been studying graphene and other 2D materials for more than a decade, and has earned a reputation as a master grower of these super-thin nanomaterials. It's a highly competitive field, but Loh "displays a rare combination of many unique skills", says Duan. He adds that Loh's training as a physical chemist allows him to bring perspectives from both fields.

He also brings insights from around the world. Loh did his PhD at the University of Oxford, UK, and his postdoctoral research at the National Institute for Materials Science in Tsukuba, Japan. He chose to return home in 1999 because, he says, "the funding situation in Singapore is very healthy" and he felt the pull of his country. "Ultimately, it's home." His work has paid off there, earning him the 2014 President's Science Award, the highest recognition given to researchers in Singapore.

Next, Loh wants to focus on translating his research into useful products that will help contribute to his homeland. "If you just focus on publishing papers, there's no value capture for Singapore," says Loh.

Already, his work on graphene and other super-thin materials has found several environmental applications, ranging from green alternatives to metal catalysts (C. Su *et al. Nature Commun.* **3**, 1298; 2012) to organic solar cells. "When you study 2D materials, they always surprise you because new properties appear," says Loh. "It's a very rich playground for intellectual discovery." ■



AMRITA CHANDRADAS FOR *NATURE*

YVONNE AI LIAN LIM

IAN TEH FOR *NATURE*

INFECTION FIGHTER

A parasite researcher in Malaysia works to improve the lives of indigenous people.

BY YAO-HUA LAW

Yvonne Ai Lian Lim has developed a reputation for her interest in human faeces. As part of her PhD project in the 1990s, Lim collected samples from the indigenous people of Peninsular Malaysia in the west of the country and checked for signs of parasitic protozoa. As she went from hut to hut, her study subjects called out, "The stool person is coming!" She made fun of her title — and that endeared her to those she was studying.

Two decades later, Lim is now a parasitologist and deputy dean in the Faculty of Medicine at the University of Malaya in Kuala Lumpur, where she continues to study parasitic helminth worms among the indigenous tribes, known collectively as the Orang Asli. She has found that a heavy rate of helminth infections persists among only certain populations, generally in places hit by poverty and poor sanitation. Working out why is a complex problem, and remains a focus of Lim's work.

She reaches into her drawer and pulls out a pouch woven from leaves. "The handiwork is excellent, it's so fine!" says Lim of the pouch, which was produced by one Orang Asli tribe. Members of that group have boosted their income by selling their craftwork, says Lim, who has connected with lawyers, economists and artists to help the Orang Asli to market their products.

Her work is also having a scientific impact. In 2016, she received a top research scientist award from the Academy of Sciences Malaysia, and co-authored a paper in *Science* (D. Ramanan *et al*. *Science* **352**, 608–612; 2016) that is in the top 1% of the most-cited papers in immunology for 2016. The study revealed a flip side to helminth worms: that a low-level infection protects people against inflammatory bowel disease by regulating the immune system and gut bacteria. The work has won Lim and her collaborators a four-year grant to study helminths as therapeutic agents. "We always think of the bad effects of the worms," says Lim. But when the worms are present in low levels, she says, they "can be angels too".

Helminth infection levels among the Orang Asli, however, are much higher and inflict a heavy toll. Lim says she will continue gathering information and pushing the issue with authorities until the problem is solved. The work has a personal dimension for Lim, who looks back on her doctoral work among the Orang Asli as a life-changing experience that keeps drawing her back to the region. She learnt from them, she says, "that life need not be so complicated, that life can be simple, that the richest people in the world are those that need the least." ■

VERONICA SANCHIS FOR *NATURE*

METAL MAESTRO

A chemist looks for new ways to make molecules shine.

BY DAVID CYRANOSKI

When Vivian Yam turns on a light, she sees a problem — and an opportunity. Much of the electricity produced globally goes into lighting, and a large portion of that energy is wasted by inefficient bulbs. But many thriftier lights contain environmentally hazardous materials such as mercury.

Yam designs organic light-emitting diodes (OLEDs) to produce cheap and environmentally friendly options. "If we can improve the performance of these devices, we can save a lot of energy," she says.

Light has been a running theme in Yam's career. The University of Hong Kong chemist has spent more than two decades creating metal-containing compounds with unique abilities to absorb and emit light. Such technologies could be used, for example, to harness solar energy, sense early signs of Alzheimer's disease in people's brains and create various types of OLED display. Her original approaches have earned her various accolades, including being the youngest person elected as a member to the Chinese Academy of Sciences.

It was a broken thermometer that helped to spark her early interest in chemistry, says Yam, who played with the spilt mercury and marvelled at how it flowed and coalesced into balls. She is still experimenting with metallic elements, but now at a molecular scale.

By manipulating the forces that hold molecules together, Yam can control their alignment. This, in turn, allows her to create larger, complex structures called supramolecular assemblies that have a unique ability to emit or absorb light. "Just the way they are

GENOME EDITOR

A South Korean researcher who helped to develop CRISPR wants to improve crops.

BY MARK ZASTROW

The détente between North and South Korea has given Jin-Soo Kim an idea for his next set of experiments. As a pioneer of gene-editing techniques, Kim has felt constrained by tight restrictions in South Korea on using biotechnology techniques to improve crop yields and improve resistance to disease. So he is thinking of doing experiments where there is less red tape, and where the government has an incentive to invest in resilient crops: North Korea.

The country has frequently struggled with famine tied to poor harvests. "Maybe this technology can be adopted very quickly to solve the current problem in North Korea, and that's exciting," he says.

Kim is no stranger to forging new paths. In the late 1990s, he had a premier position leading a team at the Samsung Biomedical Research Institute in Seoul. He was working in the nascent field of gene editing but felt trapped by the bureaucracy of a giant corporation. "I thought, 'Why work for Samsung if I can do this myself?'"

So, in 1999, he formed a company called ToolGen to develop his own genome-editing technology. It was a bold decision in a country that at the time had almost no venture-capital firms investing in biotech start-ups. The company struggled for financing in its early years, and at one point shrank from 28 employees to just 5.

Today, ToolGen is stable, with a market capitalization approaching US$1 billion. Kim, who left the company in 2005 for a post at Seoul National University, remains a shareholder and licenses his group's gene-editing technologies to it.

Among them are CRISPR–Cas9, the easy-to-use 'genetic scissors' that enable researchers to precisely cut and paste genetic material in living cells. In January 2013, Kim was one of the first to stake a claim to it, publishing a paper (S. W. Cho *et al. Nature Biotechnol.* **31**, 230–232; 2013) that showed how the tool could be used to snip at targeted spots of a DNA sequence. He holds CRISPR patents in several countries.

Kim continues to work at the frontier of gene editing, and is now jointly affiliated with South Korea's Institute of Basic Science. His lab's results include using the technology to edit the genomes of plants and to create pigs with extra muscle. And Kim is working to find therapeutic uses in humans. In 2017, he teamed up with Shoukhrat Mitalipov at the Oregon Health & Science University in Portland in an attempt to repair a mutation that can cause heart failure in humans, and then to insert the corrected gene into viable human embryos (H. Ma *et al. Nature* **548**, 413–419; 2017).

Therapeutic uses could be many years away, but Kim is optimistic that gene-edited crops could arrive sooner, with improved yields and resistance to disease. South Korea's regulation of such crops is unclear, however, and they might fall under strict guidelines for genetically modified organisms, even when no foreign genetic matter has been introduced.

Kim is considering collaborating with researchers in North Korea, which might be more receptive to editing genes in crops. By law, South Koreans are forbidden from seeking contact with North Korea. But Kim hopes that the current thaw in diplomatic relations will foster scientific cooperation. For example, students from both nations could work together in unified labs, with North Koreans returning home to train others. "I think they can do very well — they can catch up very rapidly," Kim says.

Wherever he works, Kim is intent on driving the process. "I want to see the products," he says. "I want to see at least one thing in my lifetime." ■

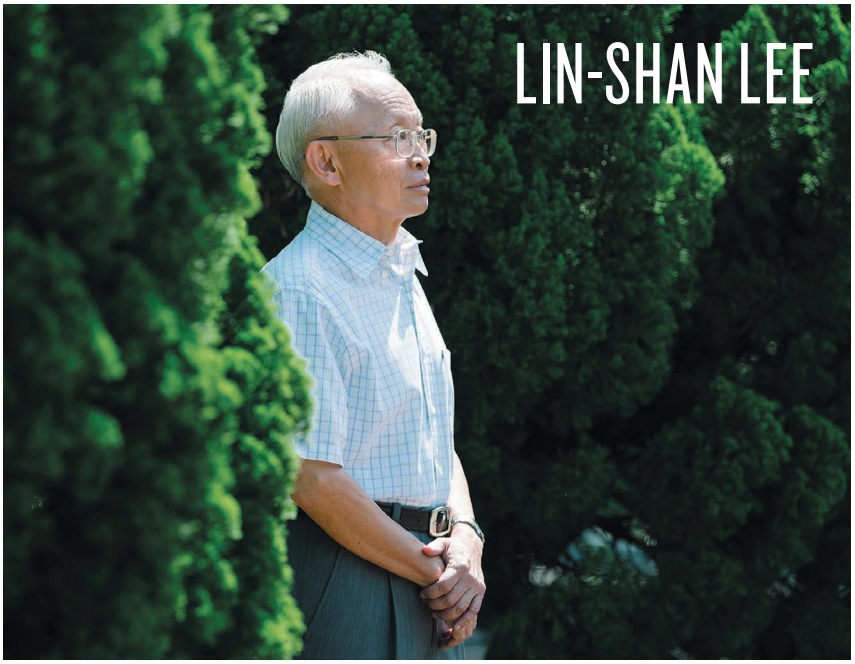
packed makes such a dramatic difference in their optical properties," she says.

Cost is a key consideration in her science. To develop phosphorescent OLEDs, known for their high efficiency in converting light into electricity, she bypassed the materials most commonly used by others: iridium and platinum. Instead, she went with gold, which is more abundant, environmentally benign, and cheaper. She eventually made the world's first gold-based phosphorescent OLEDs (V. K-M. Au *et al. J. Am. Chem. Soc.* **132**, 14273–14278; 2010).

OLEDs light up the screens of many electronic devices, from mobile phones to televisions, so Yam's work caught the attention of TCL, a firm based in neighbouring Guangdong province and one the world's largest producers of TV sets.

The company established a joint laboratory with Yam at the University of Hong Kong to develop gold-based versions of printable OLED materials, currently a hot field in the TV industry. If Yam's work pans out, her invention could light up the lives of people across the world. ■

SHIN WOONG-JAE FOR *NATURE*



JOE RUSSO FOR *NATURE*

PROGRAMMING PIONEER

A computer scientist seeks to create a spoken version of Google.

BY NICKY PHILLIPS

Lin-shan Lee spends a lot of time exploring online courses, but he isn't looking to learn a language or study ancient history. Instead, he's training algorithms to extract key words and phrases from audio and video recordings.

He and many other computer scientists say that this is one of the crucial steps towards unlocking a vast amount of knowledge that is currently difficult to search and organize. "My whole career has been motivated by identifying interesting problems which can have great impact if solved, but are very difficult," he says.

Lee, a computer engineer at the National Taiwan University in Taipei, trained as an electrical engineer, completing his PhD at Stanford University in California in the 1970s. In the United States, he also worked on satellites used for telephone communication. "When I decided to return home, all my friends said, 'you are stupid because there's no work on satellites in Taiwan'," he says.

But Lee felt that his familiarity with voice signals presented an opportunity: he traded satellites for computers, and started developing a tool that could recognize spoken Mandarin. Over more than a decade, his team built a device that was crude and slow — it took five to six seconds to recognize a single syllable — but a huge achievement. It was the first speech-recognition device for Mandarin. By 1995, their machine could transform continuous Mandarin speech into a series of Chinese characters.

"Speech recognition is a very hard problem," says Jim Glass, a researcher in speech processing at the Massachusetts Institute of Technology in Cambridge. Everyone in the world speaks slightly differently, in a continuous stream, often against a noisy background, he says. "Somehow we have to extract information from that," says Glass. "Lin-Shan is a pioneer in Mandarin spoken-language processing."

Lee is now building on that success to tackle another speech-recognition challenge: retrieving spoken content from audio and video files. His team is using machine learning to build a system in which users can search for specific words, phrases or sentences within a video. "I call this a spoken version of Google," he says. It would revolutionize multimedia sites such as YouTube. Currently, there is no way to locate a specific sentence within the video without going through the entire clip. Although there are some existing technologies that can process speech in multimedia, they are not very accurate, says Lee. "It is my research goal at the moment to develop more-accurate technology for this." ■

RNA EXPLORER

A biochemist investigates small RNA molecules that regulate gene expression.

BY MARK ZASTROW

When Narry Kim took her first faculty position in 2001, she had no funding, no students and no experience in her newly chosen field — investigating a class of recently discovered molecules called microRNAs (miRNAs).

It was a leap into the unknown for her, establishing her academic independence by leaving behind the fungi and viruses she had studied in graduate school. "It was kind of scary at the beginning. But looking back, it was a great decision," Kim says.

Now, as she sits in her well-funded lab at Seoul National University, she waxes nostalgic for those times. Her lack of money forced her to be creative, designing cheap experiments that directly attacked the key unanswered questions of how miRNAs form and what they do.

The first examples of miRNA were discovered in 1993. Much shorter than the messenger RNAs that translate genetic code from DNA to proteins, miRNAs have a role in blocking gene expression. But, as Kim began her work, the origins of the tiny stretches remained mysterious. In 2002, in her first paper as a group leader (Y. Lee *et al. EMBO J.* **21**, 4663–4670; 2002), she and her small team showed that they are generated in two steps — first in the nucleus, then with further processing in the cytoplasm into mature miRNA. She followed that up with another paper the next year detailing a major pathway through which miRNA is formed (Y. Lee *et al. Nature* **425**, 415–419; 2003).

"She and her team have made tremendous contributions," says Ling-Ling Chen, an RNA researcher at China's Shanghai Institute of Biochemistry and Cell Biology. "As a scientist, Narry is precise and dedicated."

Today, Kim is one of South Korea's most prominent researchers, and has a joint affiliation with the Institute for Basic Science, the country's premier network of blue-sky research labs. In 2009, at the age of 39, she became one of the youngest winners of the Ho-Am Prize for Medicine, often considered as South Korea's equivalent to the Nobels.

Along the way, she has also become a role model for young scientists — especially to women, who make up just 19% of South Korea's scientific workforce (see infographic, page 500).

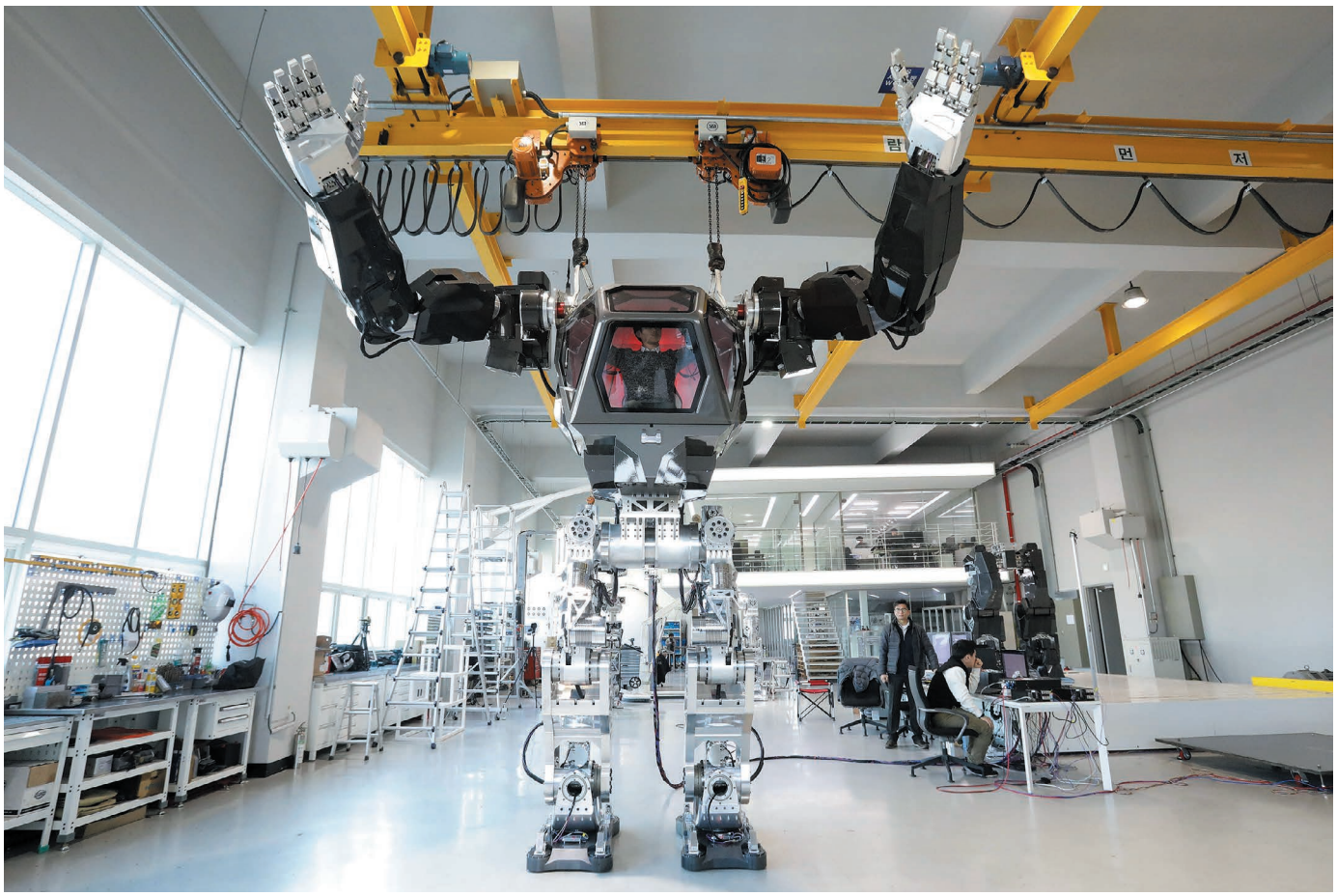
Kim recalls a moment that could have ended her career before it had a chance to take off. While on a year-and-a-half-long break from research, she started a family and moved to small towns for her husband's jobs. "There wasn't any library to read journals and no lab to work," she says. And she felt the deck was stacked against her in the country's academic culture. "That was in the late 90s, and for women, even if they excelled in science, it was very difficult to find a permanent position," she says. "I was losing hope." She decided to quit, and began studying for law school. "But it was so boring," she says. "And I wanted to go back to the lab so desperately."

Although RNA research has plenty of possible therapeutic applications, Kim says she has too little time to explore them, and firmly intends to take her mandatory retirement from academia at 65. "I have other things to do", she says. One goal is to write a novel with multiple timelines that show how science and technology change the course of human history. She sees that as a way of extending her impact beyond academia. "Storytelling is always more powerful." ■

SHIN WOONG-JAE FOR *NATURE*

COMMENT

CHUNG SUNG-JUN/GETTY



A piloted walking robot developed by robotics company Korea Future Technology in Gunpo.

Restructure science in South Korea

To build on the success of its centralized research agenda, the nation must switch to projects led by independent principal investigators, urges **Han Woong Yeom**.

South Korean science is both flourishing and floundering. In some ways, things could not be better. National spending on research and development (R&D) by industry and government was 4.24% of gross domestic product (GDP) in 2016 — the second-highest percentage for any country worldwide (Israel was first, with 4.25%). For three decades, the government has invested billions of dollars each year in high-tech industry, turning South Korean electronics companies such as Samsung, LG and SK Hynix into world leaders.

Academic research is booming, in terms of the numbers of papers published and citations: in 2016, the nation ranked 12th and 13th, respectively (see go.nature.com/2jwxzcc). Institutes and facilities are being built. In 2011, the government ▶

set up the Institute for Basic Science, a national network of research centres mimicking Germany's Max Planck Society. In 2017, it launched a world-class X-ray free-electron laser facility in Pohang[1], and in 2021, it will open a heavy-ion accelerator in Daejeon.

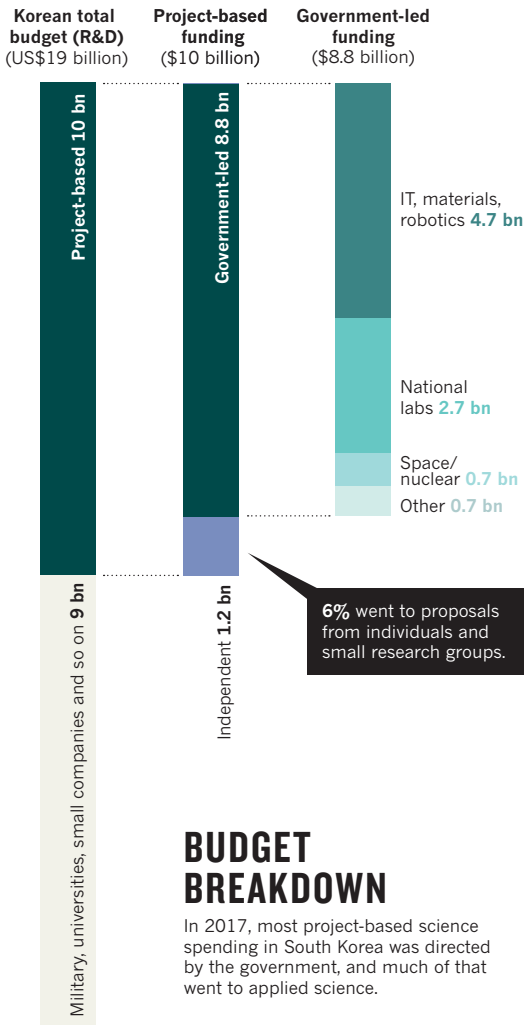
Yet many researchers are dissatisfied. In 2017, biologist Won-Kyung Ho at Seoul National University initiated a nationwide movement calling for more funding for projects proposed by individuals. In 2017, just 6% of the national research budget went to projects led by small research groups. The rest went to big, government-directed 'top-down' projects in strategic areas such as information technology, robotics, materials and biotechnology. This imbalance was arguably the biggest scientific issue discussed during the presidential election campaign last year.

Something is wrong with the South Korean R&D system, and everyone knows it. The country invests a lot but gets less and less back. Scientists feel disenfranchised by the government's opaque system for funding. The public is not seeing solutions to pressing problems, such as air pollution. Even government ministers and economists complain that all those highly cited papers are not generating enough new technologies.

But that is not so surprising. An approach that gets a nation from science infancy to adolescence might not be the best to take it to maturity. Now that South Korea's high-tech industry is world-leading, it no longer needs to be directed by the state.

The country's R&D is at a turning point. It needs a new strategy. South Korean industry must change from fast follower to first mover. Government R&D must pave the way. But I think that backing certain industries and technologies is no longer the best way to do it. A better bet is long-term support of basic science, investment in infrastructure and human capital, and addressing public issues.

As vice-chair of South Korea's Presidential Advisory Council on Science and Technology, here I outline the three main challenges and some steps taken in response. I call for urgent discussions among scientists, industrialists and policymakers to forge a more effective path.



BUDGET BREAKDOWN

In 2017, most project-based science spending in South Korea was directed by the government, and much of that went to applied science.

SOURCE: SOUTH KOREAN GOVERNMENT

THREE PROBLEMS

Bottom-up basic research is underfunded. In 2017, only US$1.2 billion of South Korea's $18 billion of government research funding went to bottom-up (independent) basic research projects by individual investigators and small groups (see 'Budget breakdown'). Eighty per cent of these projects received less than $50,000 per year, which is insufficient to fund a globally competitive study. Just 11% of grants proposed by mid-career researchers were accepted last year. This is much less than the roughly 30% acceptance rate that is standard in the United States, the European Union and Japan.

This is mainly the result of previous governments' policies, which focused on strategic areas and the development of emerging technologies, industries and markets to fuel economic growth. For example, between 1999 and 2013, $1.4 billion went to 16 projects in the 21st Century Frontier R&D Program. These ranged from nanomaterials and nanodevices to proteomics and hydrogen fuel. Over the 2010–23 period, ten projects in the Global Frontier Program run by the Ministry of Science and ICT will consume similar budgets. Topics range from flexible electronics to biomedical technology.

This approach has produced a two-tier R&D system. Little has been spent on research infrastructure in universities, and young researchers find it hard to obtain start-up grants. By contrast, some leading Chinese universities offer start-up grants of $20 million to promising researchers in areas such as condensed-matter physics — ten times more than a South Korean scientist might expect. Unsurprisingly, many talented young physicists from South Korea now work in China.

R&D is inefficient. South Korea produces many patents, but few innovations. For example, in 2016, South Korea ranked third in the world for the number of patents it filed, after the United States and Japan[2]. Yet the Netherlands, which ranked sixth and has one-quarter of South Korea's R&D budget, earned five times more in technology payments. South Korea scores low in terms of its 'innovation potential' (its ability to create and use innovations), which includes factors such as the quality of its research institutes, workforce and collaborations between universities and industry. The World Economic Forum's *Global Competitiveness Report 2017–18* lists South Korea as 26th in the world, behind New Zealand (13th), Taiwan (15th), Malaysia (23rd; see page 514) and Ireland (24th)[3].

Part of the problem is poor government leadership of R&D projects. Government-driven projects are often badly designed and can have unclear goals that fall below the standards used in industry. For example, the country's next-generation superconducting technology project began in 2001 without a clear idea of market needs, and ended up without any serious research output. None of the test products developed by the 21st Century Frontier R&D projects had a substantial market impact by 2013, when the initiative ended.

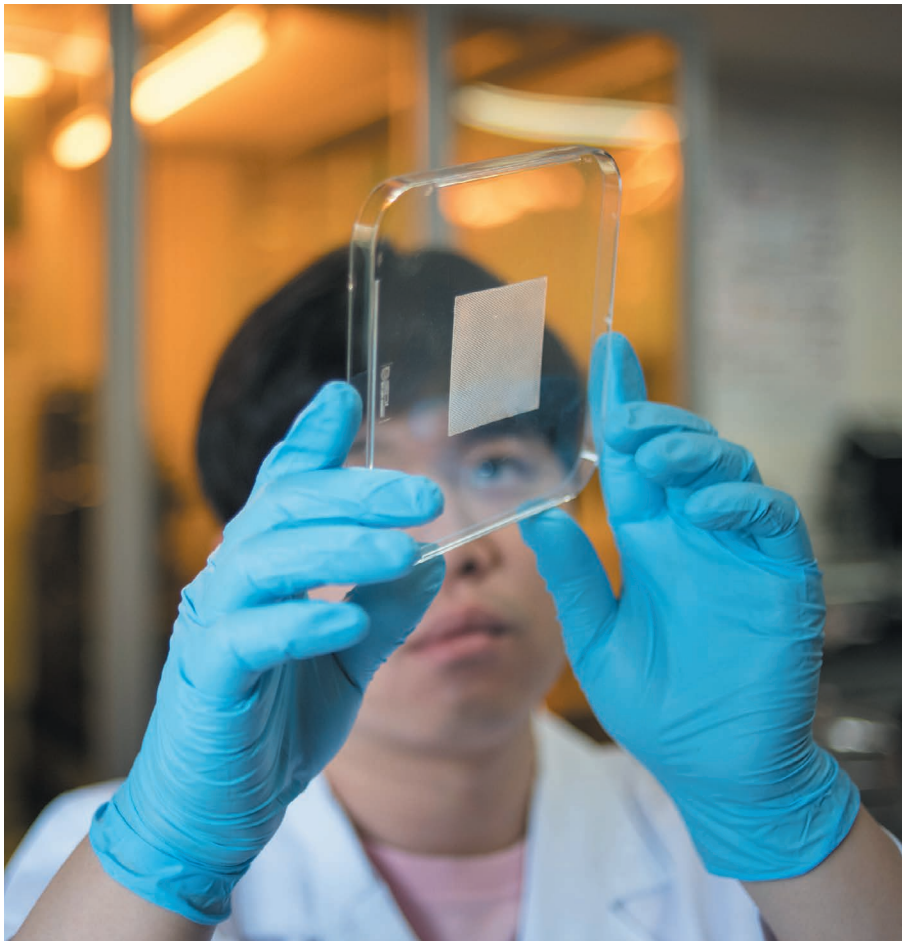
Unsurprisingly, industrialists, policymakers and the public are dissatisfied with the outcomes. Academics feel that mega-projects produce fewer papers and citations than would all the small, independent projects that could have been funded. International collaborations and the infrastructure needs of the science community are being neglected.

Public concerns are not being met. South Korea's citizens feel that they are paying for science that is not addressing their problems. Severe air pollution is the main example. South Korea's air has the highest concentration of fine particulates (PM_{10} and $PM_{2.5}$) out of the 34 countries in the Organisation for Economic Co-operation and Development: triple that of the United States and twice that of Japan. Neither the government nor the scientific community collects proper data on air quality or health impacts.

In South Korea, the proportion of people aged over 65 is increasing at the highest rate in the world. Conditions such as Alzheimer's disease have become a major social issue. Yet there is no clear national R&D strategy for this.

Several crises have further undermined

ED JONES/AFP/GETTY



An engineering student at Sung Kyun Kwan University near Seoul inspects a wet-tolerant adhesive patch inspired by the suckers found on octopus tentacles.

the public's trust in the science system. One was the magnitude-5.4 earthquake in my home city of Pohang in November 2017, which was possibly caused by a geothermal power plant injecting water into a fault zone[4]. Another was the deaths of more than 100 people — including children and their mothers — which were linked in 2011 to inhalation of toxic disinfectants that had been sold for use in home humidifiers since 2001. In both cases, scientists and the government were criticized for dodging their responsibilities.

POLICY DRIVES

The current government acknowledges all of these problems. Within two months of winning the election in May 2017, it proposed policy changes in three areas over the next five years. These are welcome, but leave many issues unaddressed.

First, the government will double the budget for bottom-up research proposals, reaching $2.2 billion in 2021. This is a good start. But universities and national labs need improvements to infrastructures, core facilities, technical capacities and research administrations. Students and postdoctoral researchers need more financial support. Addressing all of these areas could require as much funding again.

Second, more government initiatives will focus on societal issues, such as air pollution, earthquakes, chemical hazards, infection, climate change, Alzheimer's disease and renewable energy. However, this doesn't solve the inefficiency problem with government-led programmes. I am already seeing poorly designed and badly directed projects being set up, such as those on environmental issues and emerging technology. Late last year, South Korea's president, Moon Jae-in, called for improvements to the planning and organization of R&D projects, but these have yet to take effect.

"Something is wrong with the South Korean R&D system, and everyone knows it."

Third, industrial support needs reform. It is unclear which directions will most benefit the economy and where top-down involvement will be most effective. The current government is betting on the 'fourth industrial revolution' to generate commercial opportunities and quality jobs in artificial intelligence, data science, ultrafast mobile communication, the Internet of Things, smart cities, bioinformatics and so on. But artificial intelligence could take away as many jobs as it creates, for example. In my view, the concept of the fourth industrial revolution is too vague to pin taxpayers money on.

Furthermore, these three policy drives will squeeze budgets in other research areas. There is no clear idea which areas those should be.

OPEN DIALOGUE

I don't have all the answers. But the first thing that needs to happen is a frank and honest debate. Whatever policy we want to make or whichever problem we want to solve, openness and rigour will reduce the possibility of major errors.

I would like to see more scientists getting involved in policymaking in South Korea. I will seek to establish a culture that makes it routine for the government and the scientific community to hold regular, transparent, evidence-based discussions. As a first step, the Presidential Advisory Council on Science and Technology will look at South Korean R&D directions and at ways to strengthen its research institutes.

The roles of government and industry need to be reset in light of the rapidly changing global economy. Sweden, for example, has come under fire in the past for its high R&D spending and low output[5]. The rate of innovation is slowing worldwide[6], and each advance is becoming more expensive. For example, with silicon technology approaching its fundamental limit, the cost of creating computer chips that are ever smaller has exploded.

South Korean science has come a long way in the past 30 years. The challenges it faces present a great opportunity: it is time for South Korea to reposition itself to deliver the best science, innovations and solutions for its people. ■

Han Woong Yeom *is professor of physics at Pohang University of Science and Technology and director of the Center for Artificial Low Dimensional Electronic Systems at the Institute for Basic Science, Pohang, South Korea. He is vice-chair of South Korea's Presidential Advisory Council on Science and Technology.*
e-mail: yeom@postech.ac.kr

1. Kang, H.-S. *et al. Nature Photon.* **11**, 708–713 (2017).
2. Organisation for Economic Co-operation and Development. Patents Database (OECD, 2016); available at https://stats.oecd.org
3. Schwab, K. (ed.) *The Global Competitiveness Report 2017–2018* (World Economic Forum, 2018).
4. Kim, K.-H. *et al. Science* **360**, 1007–1009 (2018).
5. Bitard, P., Edquist, C. & Rickne, A. in *Small Country Innovation Systems: Globalisation, Change and Policy in Asia and Europe: Theory and Comparative Framework* (eds Edquist, C. & Hommen, L.) Ch. 7 (Edward Elgar, 2008).
6. Youn, H., Strumsky, D., Bettencourt, L. M. A. & Lobo, J. *J. R. Soc. Interface* **12**, 20150272 (2015).

MOHD SAMSUL MOHD SAID/GETTY

Improving health and well-being in Malaysia is a priority for the government.

A road map for Malaysian science

Plug the digital divide, improve health care, develop biotechnology industries and innovate for the trillion-dollar Halal economy, urges **Asma Ismail**.

Sixty years after independence, Malaysia has the third-highest gross domestic product (GDP) per person of the ten countries in the Association of Southeast Asian Nations (ASEAN). It has achieved this by shifting from producing cheap goods such as tin, rubber, cocoa, timber and rice to more lucrative commodities including oil, natural gas and palm oil. Science and technology are central to the country's economic growth strategy. For example, electronic components, petroleum products and petrochemicals, machinery and food are major exports. The government views biotechnology as a means of enhancing prosperity and wellness.

But new paths must be forged. The economy has slowed since the Asian financial crisis in 1997, as the advantages that have historically made Malaysia attractive for investment — pools of low-skilled labour, good infrastructure and tax incentives — have diminished. The nation is vulnerable to competitors such as China that produce luxury goods, and countries with lower labour costs, such as Vietnam. However, as a majority Muslim country, Malaysia is well positioned to contribute scientific innovations to premium products and services for the Halal economy.

The nation must tackle three further emerging problems — health problems, rapid urbanization and climate change. Malaysia has the highest level of obesity in southeast Asia, with nearly half of adults classified as overweight or obese. The average age of the population is rising faster than for any other country in the world. By 2050, 20% of Malaysians will be aged 60 or older, compared with 8% today[1]. By 2050, 9 out of 10 Malaysians are expected to live in cities, putting pressure on resources. Malaysia is already on average 1 °C warmer than it was in 1999; an increase of another 2 °C is predicted by 2050 (go.nature.com/2hdr6sj), along with a higher sea level and increased rainfall.

Malaysia must build on its strengths to overcome these challenges. The country is one of the most biodiverse on Earth, with almost 19,000 species[2] of animal, plant, fungus and moss as well as ancient tropical rainforests. It has more than 400 vibrant institutions of higher learning and untapped pools of talent, producers and consumers.

Malaysia spends a small proportion of its GDP on research and development (R&D), allocating the sector 1.3% in 2015, compared with 2.5% on average for developed nations[3]. A lack of national coordination is hampering R&D. Few researchers work in industry, and too little is being spent on translating R&D outputs into products and services in comparison to developed economies, such as South Korea and the United Kingdom[4]. Research is diffused across many institutions.

Malaysia needs a new national-level strategic direction for science, technology and innovation to better focus its efforts. As president of the Academy of Sciences Malaysia (ASM), an independent body mandated to be the country's thought leader for science, technology and innovation, I outline how the science and technology system can be re-engineered to help the economy and people.

MANY BODIES

The Malaysian science, technology and innovation landscape is extremely diverse. Science cuts across 23 federal ministries, and the country has 14 state governments and territories[4]. Overall, 268 entities are involved in implementing policies and initiatives. Central coordination and planning is needed to minimize duplicated efforts, and to overcome the fragmented resources and ineffective decision-making that dilute impacts. Of the total R&D budget, 70% (10.6 billion Malaysian ringgits; US$2.6 billion) was spent on applied research and 21% went to basic research — proportions in line with developed countries such as the United Kingdom and Singapore. But Malaysia falls short on investment in experimental development, spending just 9% on activities such as prototyping, piloting and scaling-up. By contrast, South Korea, Japan and Ireland spend 62%, 64% and 48% of their R&D budgets, respectively, on translating research into products and services[4].

Institutions of higher learning host 78% of Malaysian researchers; only 12% are in business enterprises and 10% are in the

government sector[4]. In South Korea, 80% of the country's researchers were in businesses in 2015. In Singapore and Thailand, it was 51%. This shows that Malaysia's R&D is not industry-driven and the lack of high-skilled talent in businesses is limiting translation of research outputs. One reason science has not percolated well into Malaysian industry is that 98.5% of businesses are small or medium-sized enterprises (SMEs). These SMEs lack skilled workers and money to invest in R&D. Three-quarters of the workforce is low-skilled, which can make integrating science a challenge. Large firms were more than three times more productive than SMEs in 2016, in terms of added value per worker, contributing 63% to the nation's GDP. SMEs should step up their productivity and contribute more to economic growth and social well-being.

Networking and collaboration could enhance SME productivity in the short term. But in the long term, Malaysian businesses need to build up and engage more scientists. This will propel them from imitating and adapting to creating.

Another problem is that companies rarely engage with communities who will use the technologies, especially the poorest. Few manufacturers want to produce products for markets that they perceive as unprofitable.

PRIORITY AREAS

In 2017, an ASM study on emerging Malaysian science and technology highlighted five technology areas for investment: biotechnology, digital technology, green technology, nanotechnology and neurotechnology (see 'Technology game-changers')[1]. These cut across disciplines and would potentially boost both industry and society.

The Halal economy is another area in which Malaysia is a leader and could contribute more scientifically, by developing products, standards and tracking systems.

Consumers spent US$2 trillion on Halal products and services globally in 2016, and 1.7 billion consumers adhere to Halal principles such as avoiding animal products and alcohol in pharmaceutical ingredients[5]. These principles are of increasing interest beyond Islamic societies. The Halal market extends from food to finance and banking, insurance, education, health care and tourism. In 2016, the Halal industry accounted for 7.5% of Malaysia's GDP, with 1,400 local enterprises exporting products worth 205 billion ringgits. The government's goal is that by 2020, the Halal industry should make up 8.7% of GDP.

Malaysian scientists are well positioned to develop quality standards and certification systems for Halal products. A range of alternative and substitute ingredients are needed for pharmaceuticals, cosmetic products and foods. Quality and safety must be built in. Food traceability is crucial to allow companies to have confidence in supply chains and to isolate affected goods quickly should a problem arise. For example, geographic information systems (GIS) and radio-frequency identification trackers can monitor products being transported from farm to user.

Malaysia is a leader in Islamic finance, having a large asset base, many institutions and a strong regulatory framework in this area. The country should encourage the development of alternative forms of financial technology through cloud computing, big-data analytics, artificial intelligence and blockchain.

PULL TOGETHER

Malaysia wants to be a strong scientific nation. Successive governments have paved the way, seeing science and technology as the backbone of development. Now the country needs a lean and well-governed R&D system, spearheaded by a central agency and centred on people.

The new government, elected in May, must start by setting up a central research management agency — similar to the US National Science Foundation — to co-ordinate, plan, monitor and evaluate the country's science, technology and innovation. That agency could be complemented by a technology commercialization agency — for example, one resembling Innovate UK — to drive innovation and commercialization.

More funding will also be needed to promote the later stages of product development. The research and commercialization agencies must forge links among industry, academia and communities.

Developing international collaborations will be important to share Malaysian strengths and values, to exchange knowledge and to spur innovation.

To be a progressive nation, Malaysia must pursue economic growth, human development and societal well-being. ■

TECHNOLOGY GAME-CHANGERS

Five priorities for Malaysian research

Investment in the following areas would best support the nation's growth[3].

Biotechnology. Agriculture has been a key sector for Malaysia's economic growth. Precision farming is set to increase yield, efficiency and market reach, and sustainable agriculture needs to be developed. Biotechnology in health care is another priority. For example, transgenic plants and genetically engineered animals might produce pharmaceutical substances more cheaply than conventional methods.

Digital technology. Malaysia has to tackle its digital divide, especially between rural and urban communities. Whereas 78% of households nationwide have Internet access, only 38% of rural households do[6,7]. Improving broadband access would enable poor and rural communities to share information and create jobs and wealth. Real-time data analysis, through the Internet of Things and cloud computing for example, could inform decision-making across sectors.

Green technology. Malaysia has pledged to reduce greenhouse-gas emissions by an ambitious 45% by 2030 as part of the Paris Agreement. Renewable-energy technologies must be advanced to reduce dependency on fossil fuel. Technologies might include fuel cells, solar and ocean thermal-energy technologies.

Nanotechnology. Real-time health monitoring using nanosensors and wearables would improve preventive health care. People could track activities, vital signs, medications and mental and physical states to facilitate early intervention. Affordable over-the-counter biomarker kits should be developed as diagnostic tools, because survival rates are higher for diseases such as cancer if they are caught early.

Neurotechnology. Neurological diseases, such as Alzheimer's and Parkinson's, are on the rise and are expected to become the second leading cause of morbidity and mortality in Malaysia, after heart disease. Incidences are expected to rise in the next decade, owing to longer lifespans and lifestyle and environmental changes. Neurogenerative treatments are needed.

Asma Ismail *is the president of the Academy of Sciences Malaysia, Kuala Lumpur.*
e-mail: asmainformm@yahoo.com

1. Academy of Sciences Malaysia. *Envisioning Malaysia 2050: A Foresight Narrative* (ASM, 2017).
2. Ministry of Natural Resources and Environment. *National Policy on Biological Diversity* (MNRE, 2016).
3. Academy of Sciences Malaysia. *The Emerging Science, Engineering and Technology (ESET) Study as part of the Envisioning Malaysia 2050: A Foresight Narrative* (ASM, 2017).
4. Academy of Sciences Malaysia. *Science Outlook 2017: Converging towards Progressive Malaysia 2050* (ASM, 2018).
5. Thomson Reuters. *Outpacing the Mainstream* (Thomson Reuters, 2017).
6. Malaysian Communication and Multimedia Commission. *Internet Users Survey 2016 Statistical Brief 20* (MCMC, 2016).
7. Malaysian Communication and Multimedia Commission. *Industry Performance Report* (MCMC, 2016).

ITAI BENIT

The Form and Function room at the Steinhardt Museum of Natural History in Tel Aviv.

NATURAL HISTORY

Israel's wild treasury

Josie Glausiusz tours Tel Aviv's stunning new museum of natural history.

A glass case crammed with a glorious jumble of stuffed animals and birds greets me as I enter the atmospheric Treasures of Biodiversity gallery at the new Steinhardt Museum of Natural History in Tel Aviv, Israel. This cabinet of curiosities contains century-old specimens collected by German ornithologist Ernst Schmitz. In 1908, Schmitz travelled to Jerusalem — then part of the Ottoman Empire — and opened a natural history museum showcasing a spectacular collection of rare mammals, butterflies and other insects, reptiles and birds' eggs, many gathered by local Bedouin people. These riches are now on display among the museum's vast holdings.

"This is the last Nile crocodile in Israel," says Tel Aviv University zoologist Tamar Dayan, chair of the museum and the driving force behind its creation. "This is the last cheetah in Israel. The leopard comes from the Judean Mountains, from a population that no longer exists. This is the last oryx in Israel, and the last bearded vulture. We have a lot of 'lasts' here." The glass case encapsulates the loss of many species — a full third of the large mammals in what is now Israel. Conveying the magnitude of this change is a key part of the museum's mission.

Steinhardt Museum of Natural History
Tel Aviv, Israel.
Preliminary opening from 2 July.

In this, it succeeds. Much of the exhibition space is inevitably allocated to displaying some of its 5.5 million skeletons and other specimens of flora and fauna, previously dispersed over the campus of the adjacent Tel Aviv University. But the causes of biodiversity loss in Israel are also front and centre. Charts in Hebrew, English and Arabic deliver the facts on climate change, habitat destruction, invasive species, pollution and over-exploitation of marine resources. A 6-metre-long interactive map of the region enables visitors to witness a century and a half of environmental impacts. For example, over-use of water has indirectly shrunk the Dead Sea, and unchecked construction has destroyed some two-thirds of Israel's coastal sand dunes since the country was established in 1948. It points, too, to the impacts of invasive species, pollution and hunting.

Some two decades in the making, the museum itself is an architectural jewel, designed by Kimmel Eshkolot Architects in Tel Aviv to resemble a treasure chest. The light-filled building — partly surrounded by an ark-like façade of wood veneer — can be navigated entirely on ramps and walkways, and is fully accessible for wheelchairs and prams. As at the Darwin Centre in London's Natural History Museum, visitors can gaze through windows beyond the taxidermy

to watch scientists studying taxonomy and systematics, biogeography and ecology, invasion biology, evolution, zooarchaeology, archaeobotany and more. Five hundred researchers will work in the museum.

The ramp system drew some criticism at the design stage from people concerned that it would encourage children to run around. Architect Michal Kimmel Eshkolot thinks that this is a good thing: public buildings, she tells me, "should be friendly to people, inviting, the opposite of intimidating".

"This is the last oryx in Israel, and the last bearded vulture. We have a lot of 'lasts' here."

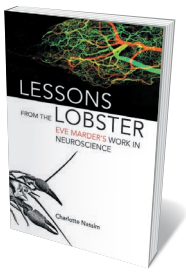
Children are just as likely to be mesmerized by the exhibits. The Form and Function room is filled with artfully mounted animals — including a caracal leaping to capture a ground-dwelling bird called a black francolin— and the skeletons of native and non-native mammals and birds. A replica velociraptor hangs from the ceiling (although, as Dayan admits, the only traces of dinosaurs found in Israel are footprints). Bugs and Beyond celebrates entomology, including cases filled with live Madagascar hissing cockroaches, giant stick insects and pinned tarantulas and butterflies of all colours and sizes. Kids can also watch a video of a dead rodent decomposing in half a minute, its fur and bones replaced by grass and a graceful toadstool.

Creative use of space is integral. Six shallow dioramas showcase the mammals, birds and insects found in six Israeli habitats: desert, sandy, aquatic, Mediterranean woodland and scrubland, the Hermon mountains and non-native, fire-prone pine plantations. A cast of a giant minke whale skeleton (in whose bones pigeons nest) hangs in an alcove open to the sky.

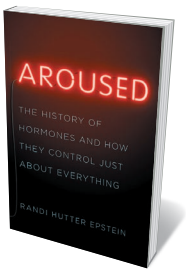
The museum's most dramatic display is also airborne. The Great Bird Migration is a swirl of avian travellers soaring high above our heads in the great entrance hall. Among them are the common crane, grey heron, greater flamingo, osprey, white stork, great white pelican and black kite. They represent some of the 500 million birds that fly over Israel in spring and autumn. Many are listed among the winged creatures that the Hebrew Scriptures' book of Leviticus prohibits the Israelites from eating. It's a reminder, perhaps, that these birds have been present in the region for thousands of years and probably much longer. And that, if we do not protect them, this is how they will end up: stuffed. ■

Josie Glausiusz *writes about science and the environment for magazines including* National Geographic, Scientific American *and* Hakai.
Twitter: @josiegz

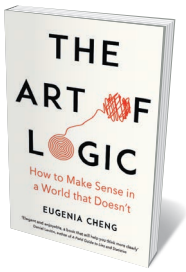
Books in brief



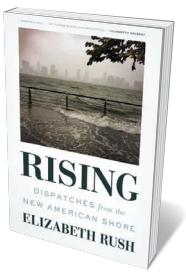
Lessons from the Lobster: Eve Marder's Work in Neuroscience
Charlotte Nassim MIT Press *(2018)*
For 40 years, neuroscientist Eve Marder has researched a tiny clutch of specialized neurons controlling the crustacean stomach — the stomatogastric ganglion. From that intense, data-centred process, she has gleaned key findings on the operation of neuronal circuits, neuronal homeostasis and neuroplasticity. Charlotte Nassim's richly detailed 'thought biography' unpeels the minutiae of lab life, revealing how Marder, "without technological fireworks or lavish funding", has illuminated areas of human neuroscience such as brain variability. A nuanced portrait of an inspired scientist at work.



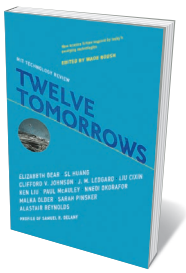
Aroused
Randi Hutter Epstein W. W. Norton *(2018)*
Hormones may be ringmasters of the bodily circus, controlling everything from sex to metabolic function, but in this invigorating history they become stars of the show. Medical journalist Randi Hutter Epstein navigates endocrinology's messy evolution through players such as neurosurgeon Harvey Cushing, who indefatigably researched the pituitary 'master gland', and driven Nobel laureate Rosalyn Yalow, who co-invented radioimmunoassay. Here, too, is the wilder side, from a testicle-swapping experiment on roosters to the animal-ovary elixirs once prescribed for menopausal women.



The Art of Logic
Eugenia Cheng Profile *(2018)*
Nothing in the world, notes mathematician Eugenia Cheng, behaves according to logic. Yet, in an era awash with conflict, exploitation, tribalism and fake news, the "illuminating precision" offered by logic is important. Cheng harnesses the power of abstraction to explore real-life phenomena such as sexism and white privilege. She walks us through the grand terrain of logic, from axioms to proofs. And she reveals how to build arguments as long chains of logical implications — a "virtuosic and masterful" skill that, combined with intelligent emotional engagement, can cut through pervasive irrationality.



Rising
Elizabeth Rush Milkweed Editions *(2018)*
This evocative exercise in lyrical reportage by Elizabeth Rush tracks sea-level rise in the here and now, by way of the disintegrating shores and salinated soils of the coastal United States. Rush journeys from the low-lying Isle de Jean Charles off Louisiana to Maine, Florida, New York and beyond, gathering stories from field biologists, climate scientists and beleaguered citizens as she goes. She touches, too, on the ten successive Atlantic storms that became hurricanes in 2017, from Franklin to Ophelia. At once a powerful group portrait of lives and communities on the brink, and a lament for lost habitats.



Twelve Tomorrows
Wade Roush (ed.) MIT Press *(2018)*
This *MIT Technology Review* anthology is a science-fictional exploration of emergent technologies and a veritable constellation of brilliant writers, among them Liu Cixin, Ken Liu, Alastair Reynolds, Elizabeth Bear and Sarah Pinsker. Ken Liu is on masterful form in 'Byzantine Empathy', a visceral narrative shaped around cryptocurrency; Pinsker's 'Escape from Caring Seasons' plays with wrist chips and drone armies; and Bear's 'Okay, Glory' features a smart home turned kidnapper. A profile of esteemed sci-fi author Samuel R. Delany is included. 'Hard' sci-fi at its best. **Barbara Kiser**

MICHAEL TEMCHINE/TWP/GETTY

Alan Bean at an exhibition of his work at the National Air and Space Museum in Washington DC in 2009.

SPACE

The artist who walked on the Moon: Alan Bean

Richard Taylor pays tribute to the Apollo astronaut who beautifully meshed science and art.

In November 1969, when I was six years old, my father pointed to the Moon and told me that a man was walking on it. I looked up and wondered what he was doing in that remote, crater-riddled land. I later learned that his name was Alan Bean, and that he was the fourth of only 12 humans so far to walk on another world. Even in that select group, he was unique: he was the only one to record what he saw on canvas and in paint. In May, he died at the age of 86.

As my interest in space travel grew, I read about what led Bean to his Apollo 12 Moon landing. Earning an aeronautical-engineering degree from the University of Texas at Austin in 1955, he soon achieved his childhood dream of becoming a Navy test pilot. His instructor was Pete Conrad, later a fellow member of the Apollo 12 mission and Moon-walker, who became his closest friend. Inspired by the "sights, sounds and smells of high performance flying machines", as Bean put it, they hatched their plan to ride the biggest flying machine of them all.

Standing 110 metres tall, the Saturn V remains the most powerful rocket ever flown. Four months before the Apollo 12 launch, one of these behemoths had carried Neil Armstrong and his crew to the first Moon landing. But whereas Armstrong took off on a sweltering summer's day, Bean, Conrad and fellow astronaut Richard Gordon sat on their rocket engulfed by a winter thunderstorm. Thirty-six seconds into their launch, the unthinkable happened. The Saturn V was struck by lightning — twice. "I looked up at the display that had all of the caution lights and there were more on than I'd ever seen in my life," Bean recalled. Seconds away from aborting the mission, he managed to reboot the affected systems. The astronauts' nervous laughter could be heard all the way to orbit.

Previous astronauts behaved with reserve. Bean gave a glimpse of the more human side of being a space explorer. Armstrong commenced his historic landing with a deadpan "See you later", descending to the Moon's surface in tension-building silence. Bean sounded like an excited tourist. His commentary seemed to touch on whatever popped into his mind: from the view ("Looks good out there, babe, looks good") to the relief of seeing his landing spot in the Ocean of Storms ("There's that crater right where it's supposed to be"), to complementing Conrad on his flying skills ("You're beautiful").

Once Bean had ignited my six-year-old imagination, I was on my way to the life scientific. I drew make-believe planets, the real Solar System, spaceships and alien cities — even how aliens might play cricket without gravity. By 1984, close to finishing my physics degree, I was — like him — grappling with competing desires to pursue science and art.

Meanwhile, the world was celebrating the 15th anniversary of the first Moon landing. Television screens were flooded with Apollo astronauts reminiscing. Seeing the blue Earth hanging like an oasis in the inky darkness filled many of them with a deep spiritual connection to the Universe. Bean, more down-to-Earth, appreciated all that the Moon lacked. "Since that time I have not complained about the weather a single time ... I've not complained about traffic," he said. "When I got back home, I'd go down to shopping centres and ... just watch the people go by and think, 'Boy, we're lucky to be here.'"

Bean's scientific legacy is fascinating. He brought back a Moon rock known as KREEP (potassium, rare-earth elements, phosphorus). Its composition led to a new model of lunar formation: the giant-impact hypothesis. Still being refined by research, this pictures the Moon forming in collisions between Earth and one or more planet-sized objects.

Bean flew once more for NASA, in 1973: he spent a record-breaking 59 days orbiting Earth as commander of the space-station mission Skylab 3. In 1981, he left the agency to work out how best to tell his story. How could he describe what it was like to hurtle home at 40,000 kilometres per hour, or to place his thumb in front of Earth and block from view everything he knew? He found his answer in painting. He mixed Moon dust into his acrylics, and used his Apollo hammer and boots to, in his words, "sculpt a textured surface unique in all of art history". (Many of Bean's works are reproduced in his 2009 book *Painting Apollo: First Artist on Another World*.)

Bean's art is important in other ways. Apollo 8 astronaut Bill Anders's stunning photograph *Earthrise*, taken from lunar orbit, is iconic. But Bean's art goes further: it adds emotion to its extraordinary scenes. Self-described as one of the more fearful astronauts, he was aware that death was always near. That comes through in his paintings. Whether we see astronauts deploying equipment, the Service Module flying across the lunar surface, or Earth peeking above the horizon, there's a feeling of being far from home — in both distance and difficulty. The loneliness in the works reminds me of Frank Hurley's photographs of Ernest Shackleton's epic 1914–17 journey to the Antarctic.

Above all, Bean's paintings serve as an antidote to that foolish idea that emerged in the 1980s: that our brains are wired to be either artistic or scientific. Inspired by his example, I went on to be a professor of both art and science. He showed that it was a simple matter. You just follow your dreams. ■

Richard Taylor *is professor of physics, psychology and art at the University of Oregon in Eugene.*
e-mail: rpt@uoregon.edu

Correspondence

One wolf shot is too many

Denmark's position as a haven for recolonizing top predators has been seriously undermined by the shooting in April of one of the first wolves to be born in the country in almost 200 years.

The Eurasian wolf (*Canis lupus lupus*) reappeared in Denmark before 2012 after at least eight individuals immigrated from Germany and Poland (see go.nature.com/2mfaxxz). The subpopulation is protected under the European Union's Habitats Directive.

The killing of the female wolf arose from controversy among hunters, farmers and politicians over the recolonization. It was widely publicized on social media. We call for greater political accountability and better management of the country's wolf population: at a minimum, there should be protected denning areas, sufficient economic compensation for sheep farmers and satellite tracking of wolves.

Without such measures, Denmark's role in setting the agenda for international agreements such as the United Nations Red List of Threatened Species could rightly be called into question.

Christian Sonne *Aarhus University, Roskilde, Denmark.*
Aage K. O. Alstrup *Aarhus University, Aarhus, Denmark.*
cs@bios.au.dk

AI threat to drug development

Artificial intelligence could help to identify more-effective candidate drugs (see *Nature* **557**, S55–S57; 2018). However, this dream held by patients, clinicians, physicians and the public-health system could become a nightmare for the pharmaceutical industry.

The development of a clinically active ingredient generally costs hundreds of millions of euros, so the compound needs to be protected by a worldwide patent for the process to be economically feasible. A patent is granted only when a compound's application can be classified as both 'new' and 'invented'. A highly effective compound thrown up by an AI algorithm could indeed be new. Whether it is 'invented', however, is debatable. This is because the inventor might be considered as either the algorithm (so not a person) or its programmer.

It could be argued that if there is a connection between the program and the compound's structure, then it is predictable by experts and so no longer inventive. Or, if the programmer can't explain how the AI algorithm found the structure, then he or she didn't invent anything. Assigning and agreeing on intellectual-property rights will be even more complex when different parts of the process behind important discoveries involve multiple contributors.

Patent law would presumably have to be adapted, for example by acknowledging the development work through a temporary ban on imitation.

Lutz Heuer *Dormagen, Germany.*
nievenheimer1@t-online.de

Young scientists aim to prioritize patients

Translational medicine helps to prevent scientific research from being wasted by focusing on the long-term benefits for patients. As early-career researchers, we want to accelerate that process — rather than waiting for senior researchers to spearhead the necessary changes.

Our international network of aspiring clinician–scientists, called Apollo, collaborates with senior scientists from the EUREKA Institute for Translational Medicine in Syracuse, Italy. We aim to learn about the goals, opportunities and challenges of translational medicine, including finding a more-efficient way to use research funding and balancing the competing interests of the different parties involved in research. We organize meetings, student workshops, mentorship programmes and an annual training programme.

We hope to develop the skills to navigate and improve the drug-discovery pipeline. Speeding up research delivery from bench to bedside will also entail changing how scientists are evaluated, as well as promoting collaboration across disciplines (R. Benedictus *et al. Nature* **538**, 453–455; 2016).

Remi Stevelink*, Gautam Kok*
Apollo Society, University Medical Center Utrecht, the Netherlands.
info@apollosociety.eu
**On behalf of the boards of Apollo Utrecht and Toronto; see www.apollosociety.eu for details.*

Cite the online date of publication

With online delivery increasingly dominating scientific publishing, most long-established journals run papers in both print and online formats — but not necessarily simultaneously. This can affect how researchers are given scientific priority. In our view, scholars from all disciplines should use the earlier, online citation date, rather than defaulting to print as the traditional record.

In our experience, the time lag between the two can be as long as 6 months. This might be crucial for annual research evaluations, for instance, when a paper is published online at the end of one year and in print the year after (see, for example, A. L. Woerman *et al. Acta Neuropathol.* **135**, 49–63; 2018, originally published online in August 2017). Timings are also key when two competing groups publish papers that report similar findings. Young scientists, in particular, need to have the publication date of their original work accurately recorded.

Using the online publication date as the primary citation would dispel such confusion. Reference lists that include dates for both publication formats and are picked up by public databases such as PubMed are a step in the right direction. Editors, authors and indexers need to work together to manage any effect on priority when publishing a paper in both formats.

Michael A. Keller *Stanford University, California, USA.*
Stanley Prusiner *Institute for Neurodegenerative Diseases, University of California, San Francisco, California, USA.*
michael.keller@stanford.edu

Scavengers need help from IPBES

Scavenger conservation is in jeopardy, which risks amplifying current threats to biodiversity and ecosystem services. In our view, ill-informed policies for managing carrion left over from farming, hunting and fisheries are largely to blame. We call for the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) to help integrate such policies more effectively with their scientific implications as it draws up its 2020–30 programme (see *Nature* http://doi.org/cq8r; 2018).

Scavengers are in decline worldwide. Vulture populations in Asia crashed after consuming livestock carcasses that contained the anti-inflammatory drug diclofenac (J. L. Oaks *et al. Nature* **427**, 630–633; 2004). Elephant and rhino carcasses poisoned by poachers are decimating African vultures and apex predators (D. Ogada *et al. Conserv. Lett.* **9**, 89–97; 2016). And changes in how discards from European fisheries are managed could adversely affect endangered seabirds (A. W. J. Bicknell *et al. J. Appl. Ecol.* **50**, 649–658; 2013).

The declines stand to compromise nutrient recycling, disease control and waste disposal, all of which contribute to public health and food security.

Patricia Mateo-Tomás *Research Unit of Biodiversity, Oviedo University, Mieres, Spain.*
Pedro P. Olea *Department of Ecology, Universidad Autónoma de Madrid, Spain.*
rktespejos@gmail.com

PLANETARY SCIENCE

Rapid formation of Mars

An analysis of meteoritic material from Mars provides an accurate timeline of the planet's early history. The results have major implications for our understanding of the processes involved in rocky-planet formation. SEE LETTER P.586

LINDA T. ELKINS-TANTON

During their formation, many rocky planets go through a phase known as a magma ocean, during which they are mostly or completely molten. Many researchers thought that the solidification of Mars's magma ocean was protracted, perhaps lasting for up to 100 million years (Myr) after the ocean's formation[1–3]. But on page 586, Bouvier *et al.*[4] show that this process was completed in less than 10 Myr. The finding suggests that habitable conditions existed on Mars up to 100 Myr before they did on Earth.

In the contest between scientific models, empirical evidence is the arbiter. More than 100 meteorites that originated on Mars have been identified on Earth, providing samples of the Martian crust. And advances in the sensitivity of instruments that measure the concentrations of individual isotopes allow the ages of these materials to be determined with high precision.

Bouvier and colleagues looked for minerals known as zircons in Martian meteoritic material. When a zircon crystallizes from its parent magma, its crystal structure allows stray uranium atoms to be trapped in the growing crystal, but rejects lead atoms. Consequently, when researchers study these minerals billions of years later, they can be confident that any lead in the crystals was produced by uranium decay and that no other sources of lead need to be considered. Furthermore, two uranium–lead decay processes (^{235}U to ^{207}Pb and ^{238}U to ^{206}Pb) can be used simultaneously to improve the precision of the results. Uranium–lead geochronology using zircons therefore yields the most precise ages of ancient geological materials that are currently possible.

The authors analysed seven hard-earned zircons and obtained ages ranging from 4,476 to 4,430 Myr. For comparison, the first solids in the gas disk around the growing young Sun, known as calcium–aluminium-rich inclusions (CAIs), formed 4,567.3 Myr ago[5]. Therefore, in the astonishingly short interval of 90 Myr, Mars grew from dust to a planet, solidified from its initial magma-ocean state and formed a crust containing zircons.

This result already shows that models predicting a protracted magma-ocean stage on Mars[1–3] cannot be correct, but Bouvier and coworkers' study yielded even finer constraints. The lutetium–hafnium decay process, ^{176}Lu to ^{176}Hf, can be used to constrain the melting history of the zircons' parent magmas, because the two isotopes behave differently during melting. The authors found that the zircons have unusually low concentrations of ^{176}Hf. This indicates that the parent magmas had lower amounts of ^{176}Lu than would be expected if they originated from the solidified products of Mars's magma ocean. To form these parent magmas, the planet must have partially melted after it had solidified.

Bouvier and colleagues' findings provide a revised timeline for the early stages of Mars's history (Fig. 1). The planet grew to approximately its current size within less than 10 Myr (and probably less than 5 Myr) of the formation of CAIs[6,7]. It then took less than 10 Myr to solidify from its initial magma-ocean phase. To put these timescales into perspective, if the Solar System were one day old, Mars would have fully formed in the first 6 minutes. About 20 Myr after the formation of CAIs, the planet partially melted to produce magmas that rose to the planet's surface; 70 Myr later, these magmas had solidified to form a zircon-containing crust.

The rapid solidification of Mars's magma ocean has important implications for our understanding of both Mars and the planet's formation of rocky planets in general. The speediness suggests that heat was easily lost from Mars, which implies that the planet's atmosphere was relatively thin. Two processes could have produced such an atmosphere: a low release of volatile gases from the magma ocean; and a stripping of the atmosphere by the young Sun[8,9]. Researchers can now constrain the extent to which such processes occur much more closely, and can apply the results to the young Earth.

The early growth and magma-ocean phase of Mars, and, by extension, of other planetary embryos, means that at least some of the planet's formation probably happened while the gas disk was still present around the young

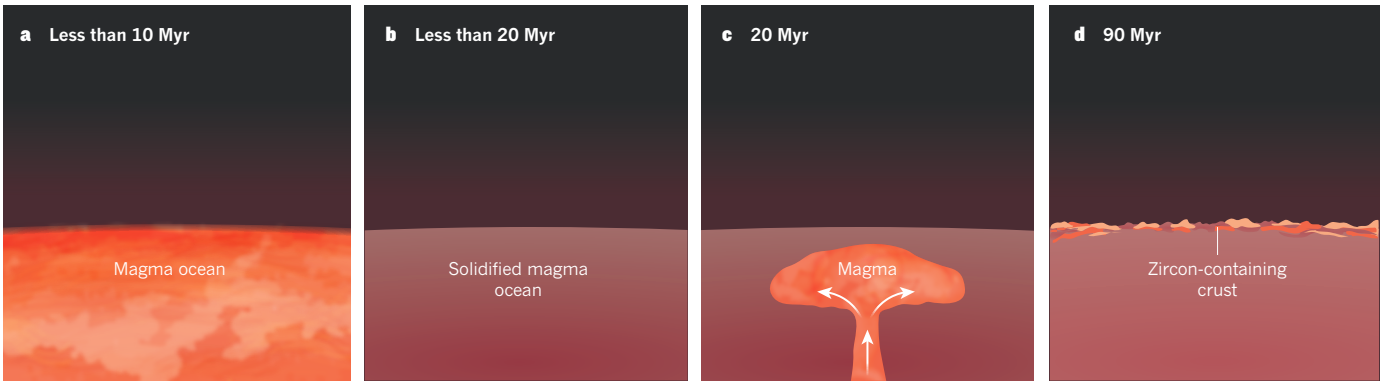


Figure 1 | The evolution of Mars. Bouvier *et al.*[4] analysed minerals known as zircons in meteoritic material from Mars, and determined a timeline for the planet's early history. The numbers represent the approximate time, in millions of years (Myr), since the formation of the first solids in the gas disk around the young Sun. **a**, After Mars had grown to approximately its current size, it existed in a magma-ocean phase, in which it was mostly or completely molten. **b**, The magma ocean solidified. **c**, The planet partially melted to produce magmas that rose (white arrows) to the planet's surface. **d**, These magmas solidified to form a zircon-containing crust.

Sun — on average, such disks exist for only a few million years[10]. Therefore, there is strong reason to think that gas in the disk would have diffused into the magma oceans on these embryos.

This diffusion process could help to answer some long-standing questions about, for example, the noble-gas content of Earth. Today, Earth releases noble gases that must have been implanted in the mantle at the time of the planet's formation. The origin of these gases has been unclear because the rocky material that built Earth contained only a small quantity of noble gases. The diffusion of noble gases from the gas disk directly into the magma ocean might solve the mystery.

Finally, Bouvier and co-workers' timeline allows the early histories of Earth and Mars to be compared directly. About 100 Myr after the formation of CAIs, Earth went through a magma-ocean phase that is thought to have been initiated by the collision of the planet with a Mars-sized body — a collision that led to the formation of the Moon[11]. Consequently, the authors' results suggest that Mars had clement conditions, and was possibly even hospitable to the formation of life, for as long as 100 Myr before such conditions existed on Earth. Mars had a head start on Earth in the planetary-evolution game. ■

Linda T. Elkins-Tanton *is at the School of Earth and Space Exploration, Arizona State University, Tempe, Arizona 85287-6004, USA. e-mail: ltelkins@asu.edu*

1. Debaille, V., Brandon, A. D., Yin, Q. Z. & Jacobsen, B. *Nature* **450**, 525–528 (2007).
2. Borg, L. E., Brennecka, G. A. & Symes, S. J. K. *Geochim. Cosmochim. Acta* **175**, 150–167 (2016).
3. Kruijer, T. S. *et al. Earth Planet. Sci. Lett.* **474**, 345–354 (2017).
4. Bouvier, L. C. *et al. Nature* **558**, 586–589 (2018).
5. Connelly, J. N. *et al. Science* **338**, 651–655 (2012).
6. Johansen, A., Mac Low, M.-M., Lacerda, P. & Bizzarro, M. *Sci. Adv.* **1**, e1500109 (2015).
7. Bollard, J. *et al. Sci. Adv.* **3**, e1700407 (2017).
8. Ikoma, M., Elkins-Tanton, L., Hamano, K. & Suckale, J. *Space Sci. Rev.* **214**, 76 (2018).
9. Erkaev, N. V. *et al. Planet. Space Sci.* **98**, 106–119 (2014).
10. Fu, R. R. *et al. Science* **346**, 1089–1092 (2014).
11. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. *Nature* **450**, 1206–1209 (2007).

MEDICAL RESEARCH

Cancer drug tackles overgrowth syndrome

Abnormal activity of the enzyme PI3K can drive cancer growth, and mutations in a PI3K subunit can sometimes lead to non-cancerous overgrowth. A cancer drug that inhibits PI3K dramatically reduces such overgrowth. SEE ARTICLE P.540

ROBERT K. SEMPLE
& BART VANHAESEBROECK

Researchers who investigate rare genetic conditions live in hope that the discovery of disease-causing mutations will lead swiftly to tailored treatments. Sadly, this is not often the case, because genetic defects usually cause impairments that are difficult or impossible to tackle using available medicines. In this issue, Venot *et al.*[1] (page 540) now provide a rare exception to this rule. In severe non-cancerous overgrowth syndromes caused by mutations in an enzyme called PI3K, they show the beneficial effects of a PI3K-inhibitor drug that was initially developed to treat cancer. Their results bring the possibility of a transformative therapy for people with overgrowth conditions one step closer.

The development of humans, from a single fertilized egg to an adult body that contains around 37 trillion cells[2] while maintaining symmetrical, paired body parts, is an astonishing feat that requires the lifelong coordination of cell division, survival and death. Growth-factor proteins can aid cellular coordination by acting on cell-surface receptors to stimulate intracellular signalling networks. These networks often include PI3K, which is essential for the regulation of growth and development by insulin and insulin-like growth-factor hormones.

Cancer arises from a flagrant breach of the rules of good cellular citizenship that are essential in multicellular organisms, and cancer cells acquire genetic abnormalities that subvert the checks and balances that constrain cell growth and migration. Mutations that activate PI3K signalling — mainly those in the gene *PIK3CA*, which encodes p110α, a catalytic subunit of PI3K — are among the most common mutations to drive solid cancers[3]. Such signalling can also be activated by mutations that inactivate the enzyme PTEN, which normally keeps PI3K activity in check. The link between overactive PI3K signalling and cancer motivated researchers to develop compounds known as PI3K inhibitors. However, the clinical impact of these drugs on cancer has been less impressive than hoped because of toxicity associated with high doses. And even when such drugs succeed in inhibiting activated PI3K, other proteins can compensate to provide alternative pathways that promote cancer[4].

In 2012, certain *PIK3CA* mutations, which had previously been linked to cancer, were reported to cause rare, non-cancerous forms of overgrowth in people[5–7]. A hallmark of these overgrowth syndromes is abnormal, excessive tissue growth that affects the body in a patchy and asymmetrical manner. This overgrowth is caused by *PIK3CA* mutations that occur after the start of embryonic development and

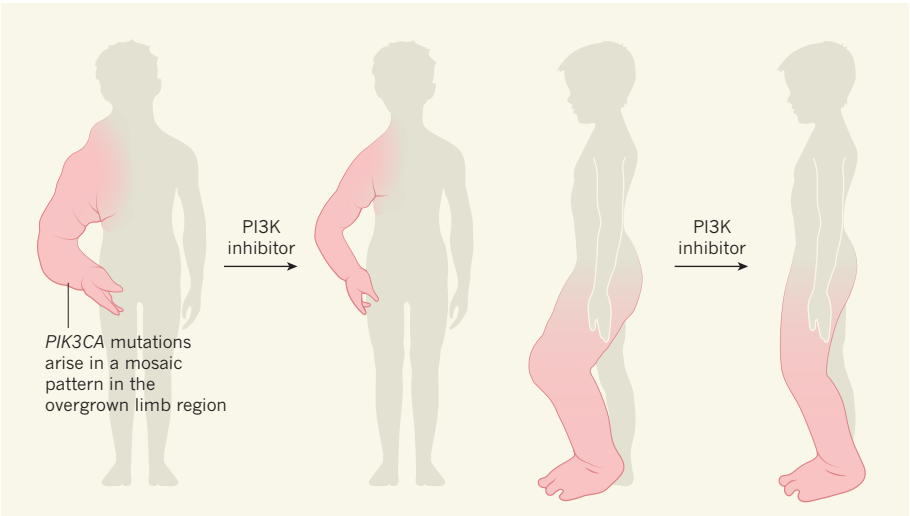


Figure 1 | People who have an overgrowth syndrome respond to treatment with a cancer drug. Venot *et al.*[1] investigated a syndrome linked to abnormal activation of the enzyme PI3K, which can result in the non-cancerous overgrowth of a variety of tissues. The authors tested whether a low dose of a PI3K inhibitor called alpelisib, developed previously as a cancer therapy, could treat people who have mutations in the gene *PIK3CA*, which encodes the catalytic subunit of PI3K. In overgrowth syndromes, these *PIK3CA* mutations arise in a mosaic patchwork pattern[5,6,8] in the region of the affected tissue (pink). Alpelisib treatment caused substantial improvements in the 19 recipients. Two examples of the decrease in overgrown tissue in patients after six months of drug treatment are shown.

only in some cells[5,6,8], which leads to cellular overgrowth in a mosaic-like pattern. The severity of the condition varies from person to person, and ranges from an isolated skin growth to a complex multisystem disorder called CLOVES syndrome[5], which comprises considerable and often widespread overgrowth that contains an abundance of fat cells and abnormal blood vessels. *PIK3CA* mutations are a common feature of many overgrowth syndromes and the term PROS (for *PIK3CA*-related overgrowth spectrum)[8] is used as a unifying description of such cases.

PROS disorders do not seem to be linked to an increase in the risk of forming the solid cancers in which *PIK3CA* mutations are most prevalent[9]. Although the reason for this is unclear, the *PIK3CA* mutations associated with such disorders usually occur in cell types of a different embryonic origin from those that develop cancer linked to *PIK3CA* mutations[9].

Severe PROS disorders can be debilitating or even life-threatening. Overgrown tissue causes compression that can lead to vascular problems or organ dysfunction. Treatments aim to reduce excess tissue by surgery or by the physical blockade of enlarged blood vessels, and other therapy options are needed urgently. The availability of targeted inhibitors of p110α that had already undergone clinical testing as treatments for cancer gave researchers hope that these might offer a new therapy for PROS disorders. Yet questions arise about the effect of prolonged patient exposure to these drugs. Would this cause side effects? Would their cells adapt to dull the effect of such treatment, as occurs in cancer[4]? And would the overgrowth be amenable to reversal by drug-based therapy?

Venot *et al.*[1] take an important step towards addressing these questions. Previous attempts to model PROS disorders in mice engineered to express *Pik3ca* containing disease-causing mutations produced excess growth in only some of the expected tissues[10]. Venot and colleagues engineered another mouse model of a PROS disorder, in which an artificial system was used to make the mice express constitutively active p110α in all tissues. These animals developed problems similar to those in people with PROS conditions, including the overgrowth of adipose, muscle and vascular tissue, and experienced a premature death caused by vascular complications. When the authors treated the mice with a PI3K inhibitor called alpelisib, an impressive, rapid and substantial decrease in the amount of overgrown tissue occurred, which prevented the premature death of the animals.

Crucially, Venot *et al*. then assessed the effects of alpelisib in 19 people with PROS disorders who had severe or life-threatening complications. In adults, the team administered the lowest dose that had been tested in trials on people with cancer (250 milligrams a day), and in children they used a dose of 50 milligrams a day. Dramatic anatomical and functional improvements occurred in all patients across many types of affected organ (Fig. 1), with some benefits noted within days of the treatment starting. The study was not randomized, blinded or subject to placebo control, yet these striking initial results suggest that this outcome is likely to have clinical importance. Resistance to alpelisib was not observed, and the drug was well tolerated by the recipients.

A predicted side effect of PI3K inhibition is a high blood glucose level, caused by interference with the PI3K-mediated metabolic effects of insulin. However, blood-glucose elevation occurred in only three people, in whom the elevation was modest. In children, the drug did not have an effect on normal growth, which suggests that overgrown tissue can be targeted without harmfully blocking PI3K-dependent childhood growth. Further systematic clinical studies are now needed, and ethics committees will have to assess whether it could be justified to include a placebo in trials on patients who are severely affected.

The study of PI3K inhibition as a treatment for PROS disorders might also offer something in return towards the design of cancer therapies. The aim of PI3K-inhibitor therapy in PROS conditions would be to suppress disease-causing levels of PI3K signalling, while minimizing any side effects during the long-term, and probably lifelong, treatment. By contrast, the conventional approach of cancer therapy involves identifying differences between healthy and cancerous cells, and then hitting the cancer-specific characteristics as hard as possible to induce the death of cancer cells. In clinical trials for cancer, PI3K inhibitors are usually studied at the maximum tolerated dose. Yet whether this makes sense is unclear, given that the activity level of mutated p110α in cancer, which is low, is probably similar to the activity level of the same subunit in PROS disorders. Could a low dose of PI3K inhibitor be beneficial in treating a cancer linked to a *PIK3CA* mutation? This might abolish any PI3K activity that is above the usual level without completely blocking PI3K signalling, as would occur with a high dose of inhibitor. This could be tested, for example, as a strategy for preventing types of cancer in which PI3K activation or PTEN inactivation is an early event[11,12], or for preventing PTEN hamartoma tumour syndrome in people who carry a mutation in PTEN and are therefore prone to cancer[13].

Low-dose PI3K inhibition might also be used as an option, following conventional treatments such as chemotherapy or surgery, for slowing the evolution of cancer and its adaptation to selective pressures[14]. And long-term, low-dose PI3K inhibition could offer further benefits — for example, it increases the metabolic health of obese mice and rhesus monkeys[15], and sustained blockade of the PI3K pathway can slow ageing in animal models[16].

Perhaps it is time to target abnormal signalling in cancer with a lighter touch, which could enable the use of combination therapies that are currently precluded for reasons of toxicity. After all, there is no need to use a hammer to kill a fly, and this principle might also apply to treating cancer. ■

Robert K. Semple *is in the Centre for Cardiovascular Science, University of Edinburgh, Edinburgh EH16 4TJ, UK.*
Bart Vanhaesebroeck *is in the UCL Cancer Institute, University College London, London WC1E 6BT, UK.*
e-mails: rsemple@exseed.ed.ac.uk; bart.vanh@ucl.ac.uk

1. Venot, Q. *et al. Nature* **558**, 540–546 (2018).
2. Bianconi, E. *et al. Ann. Hum. Biol.* **40**, 463–471 (2013).
3. Samuels, Y. *et al. Science* **304**, 554 (2004).
4. Rozengurt, E., Soares, H. P. & Sinnet-Smith, J. *Mol. Cancer Ther.* **13**, 2477–2488 (2014).
5. Kurek, K. C. *et al. Am. J. Hum. Genet.* **90**, 1108–1115 (2012).
6. Lindhurst, M. J. *et al. Nature Genet.* **44**, 928–933 (2012).
7. Rivière, J.-B. *et al. Nature Genet.* **44**, 934–940 (2012).
8. Keppler-Noreuil, K. M. *et al. Am. J. Med. Genet. A* **167**, 287–295 (2015).
9. De Santis, M. C. *et al. Cancers* **9**, 30 (2016).
10. Castillo, S. D. *et al. Sci. Transl. Med.* **8**, 332ra43 (2016).
11. Gerstung, M. *et al.* Preprint at bioRxiv https://doi.org/10.1101/161562 (2017).
12. Jamal-Hanjani, M. *et al. N. Engl. J. Med.* **376**, 2109–2121 (2017).
13. Yehia, L. & Eng, C. *Endocr. Relat. Cancer* https://doi.org/10.1530/ERC-18-0162 (2018).
14. Sansregret, L., Vanhaesebroeck, B. & Swanton, C. *Nature Rev. Clin. Oncol.* **15**, 139–150 (2018).
15. Ortega-Molina, A. *et al. Cell Metab.* **21**, 558–570 (2015).
16. Bettedi, L. & Foukas, L. C. *Biogerontology* **18**, 913–929 (2017).

B.V. declares competing financial interests.
See go.nature.com/2m8gq1t for details.

This article was published online on 13 June 2018.

QUANTUM NANOSCIENCE

Orbital insight from an upright molecule

A molecule standing on a metal surface has been found to emit electrons in the presence of an applied electric field. The emitted electrons produce an interference pattern reminiscent of a classic physics experiment. SEE LETTER P.573

THOMAS GREBER

Quantum systems are described by wavefunctions, which have an amplitude and a phase: the square of the amplitude describes the probability of finding a particle in a given region of space-time, whereas the phase describes the sign (plus or minus) of the wavefunction. The ability to control the phase of systems of electrons would open up opportunities for the development of quantum devices, and the first step in achieving such control is to ascertain what the phase is in the first place. Unfortunately, the phase of an object's wavefunction is not directly observable. It is, however, possible to work out the relative phase by observing interference patterns formed from the superposition (summation) of coherent electron waves (those between which there is a constant phase difference), by borrowing schemes from classic experiments that observed interference patterns in light, such as Thomas Young's 'double-slit' experiment[1] or Dennis Gabor's demonstration of holography[2]. Writing in this issue (page 573), Esat *et al.*[3] report a tabletop experiment that allows the phase of a molecular orbital to be determined from an interference pattern that arises as a result of electron emission from the molecule concerned.

Esat and colleagues began their investigation by assembling a molecule on a silver surface, using the tip of a scanning tunnelling microscope at cryogenic temperatures (5 kelvin) to manipulate atoms and molecules with sub-nanometre precision. More specifically, they attached two silver atoms to one end of a flat pigment molecule known as 3,4,9,10-perylene-tetracarboxylic dianhydride (PTCDA) that was lying on the surface. They then lifted up the metal–molecule complex using the tip of their microscope, so that it stood upright on the surface (Fig. 1).

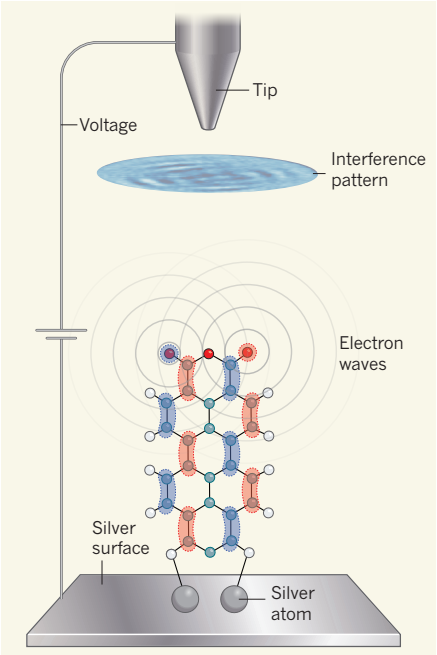


Figure 1 | A molecular electron emitter. Esat *et al.*[3] prepared a metal–molecule complex that stands upright on a silver surface. When the authors applied a voltage between the surface and the tip of a scanning tunnelling microscope, the molecule emitted electrons one at a time, producing electron waves. The spatial distribution of the resulting current contains patterns caused by interference between the electron waves. By analysing the patterns, the authors obtained information about the relative phase (plus or minus) of the region of the molecular orbital from which the electrons were emitted (the orbital is shown in blue and red; the colours represent the two relative phases).

The authors find that the complex is stable in this upright position — which might seem surprising to those in the know, because organic molecules preferentially lie flat on metallic surfaces. It is not known which conformation of Esat and colleagues' complex (flat or upright) is the more stable. However, their experimental finding casts light on how molecules such as PTCDA can be stacked on metal surfaces, knowledge of which is essential for constructing nanoscale devices in which molecules are in electrical contact with metals.

Erecting the molecule into this upright conformation allows it to perform a peculiar new function: it can emit electrons in the presence of an electric field. When the authors positioned the microscope tip 7 nanometres above the standing molecule and applied a voltage of about 25 volts, they detected an electron current of 100 picoamps (1 pA is 10^{-12} amps). Almost all of the electrons in the current pass across the sharp peak formed by the standing molecule. The electric field at the molecule's apexes is much greater than it would be between flat electrodes, because it is enhanced by the curvature of the molecule. Esat and co-workers show that the field enhancement is sufficiently high to allow electrons on the molecule to 'tunnel' into the surrounding vacuum, as measured in the field-emission current.

The authors report that the electrons undergo a two-step tunnelling process to pass from the silver metal surface to the vacuum. First, a single electron tunnels from the surface into the lowest unoccupied molecular orbital (LUMO), where it adopts the orbital's phase. In the second step, the electron is emitted at the edges of the molecule. The spatial distribution of the emitted current contains patterns caused by the interference of each electron with itself. The existence of these features indicates that the emitted electrons 'remember' the phase adopted from the part of the LUMO from which they were emitted — the

patterns wouldn't form unless the emissions had retained the orbital's phase.

Esat and colleagues' experiment is reminiscent of Young's double-slit experiment[1], in which the patterns formed by the interference of light proved that light is a wave. But, in contrast to Young's experiment, the emission patterns observed by Esat *et al.* can be explained only if the electron wavefunction has a different sign depending on whether it is emitted from the top right or top left corners of the molecule. The relative phases of the electrons emitted from different sites of the molecule can thus be worked out from the spatial distribution patterns of the emission current.

The authors used an established method for moving atoms and molecules[4] to produce their device. A complementary approach has previously been reported[5] in which electrons are coherently emitted from carbon nanotubes. The physics underpinning the emission process is the same in both systems, but the approaches used to realize it are completely different: Esat and colleagues' method can be thought of as a 'bottom-up' approach, in which the emitter is constructed from scratch, whereas the nanotube method was a 'top-down' approach in which nanotubes were painstakingly processed to allow the interference patterns to be observed and studied. The structures of Esat and colleagues' emitters are therefore much more precisely defined and reproducible.

The emission of electrons from a molecular device could, in principle, be triggered and steered using a laser, as was recently demonstrated for larger emitters[6]. This would require the stability of the molecular emitters to be improved, but would be another step towards the development of phase control. Molecular emitters might eventually find applications in devices such as electron microscopes, detectors that identify the phase or spin of electrons, or even quantum computers. ■

Thomas Greber *is at the Physik-Institut, University of Zurich, 8057 Zurich, Switzerland.*
e-mail: greber@physik.uzh.ch

1. Young, T. *Phil. Trans. R. Soc. Lond.* **92**, 12–48 (1802).
2. Gabor, D. *Nature* **161**, 777–778 (1948).
3. Esat, T., Friedrich, N., Tautz, F. S. & Temirov, R. *Nature* **558**, 573–576 (2018).
4. Eigler, D. M. & Schweizer, E. K. *Nature* **344**, 524–526 (1990).
5. Oshima, C. *et al. Rev. Lett.* **88**, 038301 (2002).
6. Yanagisawa, H. *et al. Sci. Rep.* **7**, 12661 (2017).

MEDICAL RESEARCH

Weighing in on weight loss linked to cancer

Weight loss and tissue wasting often occur in pancreatic cancer. Analyses of human and mouse data reveal a mechanism behind these events, and raise the question of whether tissue wasting affects cancer survival rates. SEE LETTER P.600

J. MATTHIAS LÖHR

An early symptom of pancreatic cancer is a profound loss in weight that can precede disease diagnosis by months[1]. Weight loss also occurs in other types of cancer, and is often associated with severe illness and a reduced quality of life. On page 600, Danai *et al.*[2] report an analysis of pancreatic cancer, using mouse models and clinical data, that illuminates the consequences of weight loss for cancer outcomes.

Cachexia, the term used to describe the cancer-linked symptom of severe weight loss, has been recognized since at least the time of the ancient Greek physician Hippocrates. It is often a hallmark of cancers originating in the gut system[3], and might manifest in changes such as loss of fat (adipose) tissue or skeletal-muscle wasting, which could arise if the body is using up the nutrient stores in such tissues. Cachexia is particularly common in people who have a type of cancer called pancreatic ductal adenocarcinoma. The mechanisms driving cachexia are not the same in all tumours[4], but whether there are different types of cachexia depending on the tumour type or the stage of the cancer at which weight loss occurs remains to be determined.

Two key mechanisms[4] thought to drive cachexia are the breakdown of molecules in a process called catabolism, and inflammation, which is controlled by the body's immune system. The pancreas secretes digestive enzymes that break down complex, calorie-rich food to provide the components needed for tissue growth and maintenance[5]; this catabolism-supporting function is known as its exocrine role. Exocrine-system impairment causes malnutrition that can lead to life-threatening tissue wasting. However, the degree to which pancreatic exocrine-system abnormalities contribute to human cachexia was unknown.

Human pancreatic cancer often occurs in a region of the organ that can obstruct the main pancreatic duct, hampering enzyme release. This can lead to a situation termed pancreatic exocrine insufficiency, which results in nutrient-absorption deficiencies and weight loss[6]. Cachexia in humans can be exacerbated if deficiencies occur in essential nutrients[7,8], for example long-chain fatty acids and vitamin D, whose uptake is facilitated by pancreatic enzymes such as lipase. The administration of fatty acids increases skeletal-muscle mass in people with pancreatic cancer, particularly when this supplementation is combined with pancreatic enzymes[9].

"This striking result indicates that cachexia does not drive cancer-associated mortality."

The research group that conducted the current study had previously[10] observed that tissue breakdown occurs before pancreatic-cancer diagnosis in humans and before the development of early-stage pancreatic cancer in mice. To continue their investigation, Danai and colleagues studied pancreatic cancer using genetically engineered mouse models of the condition[11,12], and they used transplantation experiments to test whether tumour location affects wasting. They found that if pancreatic-tumour cells were transplanted into mice beneath the skin surface, adipose-tissue wasting did not occur, whereas wasting did occur if the cells were transplanted into the pancreas. This finding indicates that some aspect of the pancreatic environment has a key role in this phenomenon, and is consistent with the results of a previous study[11]. However, that study also found that tissue wasting was promoted when tumour cells were introduced into the body cavity, suggesting that tumour presence at a non-pancreatic site can also trigger this phenomenon.

Danai and colleagues' metabolic investigations revealed that mice with pancreatic tumours used less oxygen and produced less carbon dioxide than did control mice lacking tumours. This suggested that the presence of the cancer might be linked to a decrease in the processes involved in food breakdown and nutrient adsorption. To investigate how the pancreatic-tumour environment might cause this early metabolic change and weight loss, the authors tested whether pancreatic exocrine insufficiency was responsible, given that this can occur in human pancreatic cancer[6]. When Danai and colleagues gave the mice pancreatic enzymes, the level of adipose-tissue wasting decreased, suggesting that pancreatic exocrine insufficiency has a causal role in cachexia (Fig. 1).

Danai and colleagues found that, although pancreatic-enzyme supplementation could limit the tissue wasting, the animals' survival rate did not improve. This striking

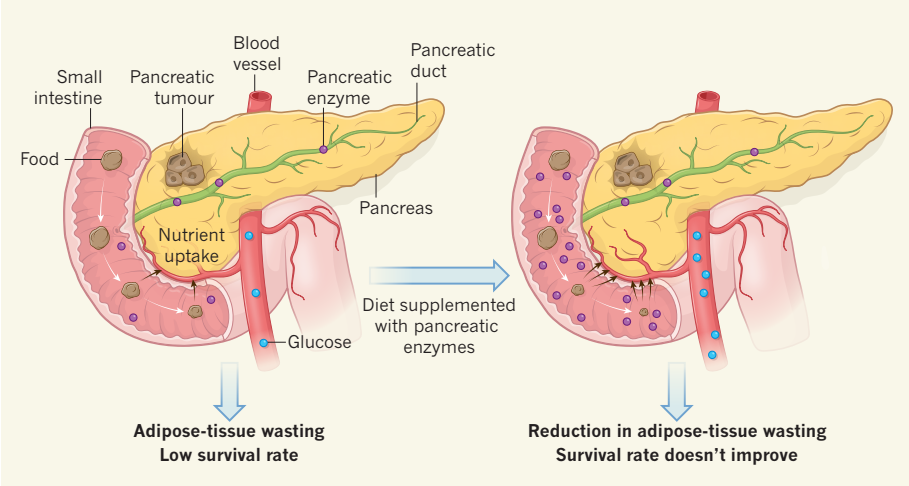


Figure 1 | Tissue wasting in pancreatic cancer. The pancreas secretes enzymes into the small intestine that aid the breakdown of food, enabling nutrients to be absorbed (black arrows). Abnormalities in this process can lead to weight loss and tissue wasting. Danai *et al.*[2] investigated the cause and consequences of the weight loss and adipose-tissue wasting that often occur early in pancreatic cancer. They observed that mice with pancreatic tumours had lower blood glucose levels, increased levels of adipose-tissue wasting and a decreased survival rate compared with control mice that did not have a pancreatic tumour. The authors tested whether feeding the mice pancreatic enzymes would result in any improvements, and found that it decreased adipose-tissue wasting and increased blood glucose. However, these changes did not increase the animals' survival rate, providing insights into the debate about whether weight loss is linked to cancer mortality.

result indicates that cachexia does not drive cancer-associated mortality. The result is also consistent with previous clinical evidence[6] that pancreatic-enzyme supplements do not improve survival in pancreatic cancer. Moreover, when Danai *et al.* analysed clinical data to assess adipose-tissue wasting in 782 people with pancreatic cancer, they found that wasting did not correlate with poorer survival rates. However, it was previously reported[13] that the loss of skeletal muscle and adipose tissue is linked to worse cancer survival rates, so Danai and colleagues' results call into question the idea that cachexia affects survival.

As well as regulating exocrine function, the pancreas has endocrine functions — it produces hormones that regulate metabolism. A key component of the endocrine system produced by the pancreas is the hormone insulin. Insulin facilitates glucose uptake into cells, and its absence can cause diabetes. Diabetes can sometimes precede pancreatic-cancer diagnosis by a year or two, and might be a red flag of trouble ahead[14]. Moreover, abnormal glucose metabolism might contribute to adipose- and skeletal-tissue wasting[15], and diabetes can cause exocrine insufficiency[16].

Danai and colleagues observed lower insulin and glucose levels in the blood of their model mice compared with the levels in control mice, and this decrease in insulin and glucose might lead to increased breakdown of stored fats, which could, in turn, increase the level of tissue wasting. This potential connection between the endocrine and exocrine systems and weight loss is supported by studies in the fruit fly *Drosophila melanogaster*[17].

Much remains to be understood about the role of the exocrine and endocrine systems in pancreatic cancer. One way to address this might be to perform detailed gene- and protein-expression analyses to determine the signalling crosstalk between transplanted cancer cells and the surrounding healthy pancreas in the mouse model used by the authors. Another potential avenue of research would be to investigate pancreatic exocrine insufficiency at the time of cancer diagnosis, especially in people whose tumours do not block the main pancreatic duct.

The authors did not investigate the role of inflammation in cancer-associated weight loss, but this is tricky to investigate because pancreatic tumours are associated with immunosuppression caused by factors such as the protein TGF-β. Inhibiting TGF-β reduced cachexia in a mouse model of pancreatic cancer[18], and there is circumstantial evidence that low-level inflammation contributes to pancreatic-exocrine insufficiency[19]. These observations provide tantalizing hints that inflammation warrants further investigation in this context.

It is worth considering whether other mechanisms might contribute to cachexia. For example, appetite loss might in turn reduce enzyme output, so dietary intake could be another key factor. As work such as that of Danai and colleagues improves our understanding of cachexia, the condition comes into focus as a distinct entity, rather than merely an early symptom of cancer. A goal for future research should be to delineate the interactions between exocrine and endocrine function and inflammation in cachexia. Although Danai and colleagues' results cast doubt on whether

50 Years Ago

The resignation of two matrons within a short space of time suggests that discontent among hospital staff is on the increase … matrons simply do not wield today the power they used to. Together with senior nurses they are assuming more and more responsibility, but their opinions are not being taken into account. It would not be a gross exaggeration to say that the concept of an all-powerful, dictatorial matron is fast disappearing, and this is perhaps no bad thing, because no individual can successfully carry the burden of running a hospital. But the answer does not lie in the appointment of honorary members of hospital committees who may be highly capable managers and administrators in their own right, but who have no real knowledge of the problems of nursing staff. What seems to be happening is that these "amateurs" … are overriding the people with professional knowledge.

From *Nature* 29 June 1968

100 Years Ago

Every boy and girl at school who "does science" now learns that metric units are the universal medium of scientific expression, and is practised in their use … A boy goes home at the end of term and tells his father that he has been doing science, weighing in grams, measuring lengths in centimetres, pressures in millimetres of mercury, and temperatures in degrees centigrade. Surely the most natural remark for any naturally minded parent to make is that his boy need not pay any attention to that, because, if it had any bearing at all upon practical life, he would certainly have been taught to use pounds or grains, inches, and Fahrenheit degrees, and not the outlandish things that nobody uses after he has left school.

From *Nature* 27 June 1918

cachexia affects survival in cancer, if progress could be made to stop tissue wasting, it would substantially alleviate the disease burden for patients. ■

J. Matthias Löhr *is in the Department of Cancer Medicine, Karolinska University Hospital, and in the Department of Clinical Intervention and Technology, Karolinska Institutet, 141 86 Stockholm, Sweden.*
e-mail: matthias.lohr@ki.se

1. DiMagno, E. P. *Ann. Oncol.* **10** (Suppl. 4), S140–S142 (1999).
2. Danai, L. V. *et al. Nature* **558**, 600–604 (2018).
3. Tisdale, M. J. *Nature Rev. Cancer* **2**, 862–871 (2002).
4. Fearon, K., Arends, J. & Baracos, V. *Nature Rev. Clin. Oncol.* **10**, 90–99 (2013).
5. Milton, K. *J. Nutr.* **133**, 3886S–3892S (2003).
6. Vujasinovic, M., Valente, R., Del Chiaro, M., Permert, J. & Löhr, J. M. *Nutrients* **9**, 183 (2017).
7. Murphy, R. A. *et al. Lipids* **47**, 363–369 (2012).
8. Dev, R. *et al. Oncologist* **16**, 1637–1641 (2011).
9. Abe, K. *et al. Anticancer Res.* **38**, 2369–2375 (2018).
10. Mayers, J. R. *et al. Nature Med.* **20**, 1193–1198 (2014).
11. Michaelis, K. A. *et al. J. Cachexia Sarcopenia Muscle* **8**, 824–838 (2017).
12. Hingorani, S. R. *et al. Cancer Cell* **4**, 437–450 (2003).
13. Di Sebastiano, K. M. *et al. Br. J. Nutr.* **109**, 302–312 (2013).
14. Gupta, S. *et al. Clin. Gastroenterol. Hepatol.* **4**, 1366–1372 (2006).
15. Kalyani, R. R., Corriere, M. & Ferrucci, L. *Lancet Diabetes Endocrinol.* **2**, 819–829 (2014).
16. Hardt, P. D. *et al. Pancreatology* **3**, 395–402 (2003).
17. Wagner, E. F. & Petruzzelli, M. *Nature* **521**, 430–431 (2015).
18. Greco, S. H. *et al. PLoS ONE* **10**, e0132786 (2015).
19. Löhr, J. M., Panic, N., Vujasinovic, M. & Verbeke, C. S. *J. Intern. Med.* **283**, 446–460 (2018).

The author declares competing financial and other interests. See go.nature.com/2sp7yeo for details.

This article was published online on 20 June 2018.

In retrospect

Twenty years of network science

The idea that everyone in the world is connected to everyone else by just six degrees of separation was explained by the 'small-world' network model 20 years ago. What seemed to be a niche finding turned out to have huge consequences.

ALESSANDRO VESPIGNANI

In 1998, Watts and Strogatz[1] introduced the 'small-world' model of networks, which describes the clustering and short separations of nodes found in many real-life networks. I still vividly remember the discussion I had with fellow statistical physicists at the time: the model was seen as sort of interesting, but seemed to be merely an exotic departure from the regular, lattice-like network structures we were used to. But the more the paper was assimilated by scientists from different fields, the more it became clear that it had deep implications for our understanding of dynamic behaviour and phase transitions in real-world phenomena ranging from contagion processes to information diffusion. It soon became apparent that the paper had ushered in a new era of research that would lead to the establishment of network science as a multidisciplinary field.

Before Watts and Strogatz published their paper, the archetypical network-generation algorithms were based on construction processes such as those described by the Erdös–Rényi model[2]. These processes are characterized by a lack of knowledge of the principles that guide the creation of connections (edges) between nodes in networks, and make the simple assumption that pairs of nodes can be connected at random with a given connection probability. Such a process generates random networks, in which the average path length between any two nodes in the network — measured as the smallest number of edges needed to connect the nodes — scales as the logarithm of the total number of nodes. In other words, randomness is sufficient to explain the small-world phenomenon popularized as 'six degrees of separation'[3,4]: the idea that everyone in the world is connected to everyone else through a chain of, at most, six mutual acquaintances.

However, random construction fell short of capturing the local cliquishness of nodes observed in real-world networks. Cliquishness is measured quantitatively by the clustering coefficient of a node, which is defined as the ratio of the number of links between a node's neighbours and the maximum number of such links. In real-world networks, node clustering is clearly exemplified by the axiom 'the friends of my friends are my friends': the probability of three people being friends with each other in a social network, for example, is generally much higher than would be predicted by a model network constructed using the simple, stochastic process.

To overcome the dichotomy between randomness and cliquishness, Watts and Strogatz proposed a model whose starting point is a regular network that has a large clustering coefficient. Stochasticity is then introduced by allowing links to be rewired at random between nodes, with a fixed probability of rewiring (p) for all links. By tuning p, the model effectively interpolates between a regular lattice ($p \to 0$) and a completely random network ($p \to 1$).

At very small p values, the resulting network is a regular lattice and therefore has a high clustering coefficient. However, even at small p, short cuts appear between distant nodes in the lattice, dramatically reducing the average shortest path length (Fig. 1). Watts and Strogatz showed that, depending on the number of nodes[5], it is possible to find networks that have a large clustering coefficient and short average distances between nodes for a broad range of p values, thus reconciling the small-world phenomenon with network cliquishness.

Watts and Strogatz's model was initially regarded simply as the explanation for six degrees of separation. But possibly its most important impact was to pave the way for

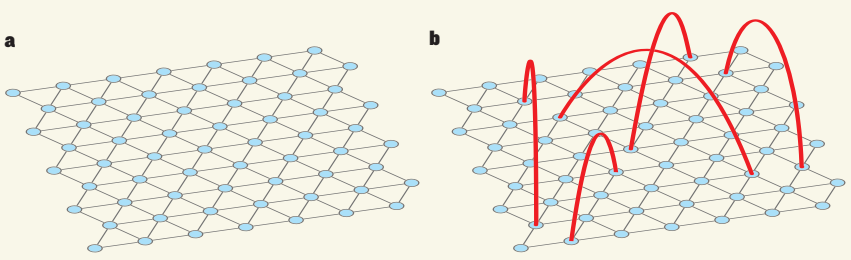


Figure 1 | The small-world network model. In 1998, Watts and Strogatz[1] described a model that helps to explain the structures of networks in the real world. **a**, They started with a regular network, depicted here as nodes connected in a triangular lattice in which each node is connected to six other nodes. **b**, They then allowed links between nodes to be rewired at random, with a fixed probability of rewiring for all links. As the probability increases, an increasing number of short cuts (red lines) connect distant nodes in the network. This generates the small-world effect: all nodes in the network can be connected by passing along a small number of links between nodes, but neighbouring nodes are connected to one another, forming clustered cliques. (Adapted from Samay/Vespignani.)

studies of the effect of network structure on a wide range of dynamic phenomena. Another paper was also pivotal: in 1999, Barabási and Albert proposed the 'preferential-attachment' network model[6], which highlighted that the probability distribution describing the number of connections that form between nodes in real-world networks is often characterized by 'heavy-tailed' distributions, instead of the Poisson distribution predicted by random networks. The broad spectrum of emergent behaviour and phase transitions encapsulated in networks that have clustered connectedness (as in Watts and Strogatz's model) and heterogeneous connectedness (as in the preferential-attachment model) attracted the attention of scientists from many fields.

A string of discoveries followed, highlighting how the complex structure of such networks underpins real-world systems, with implications for network robustness, the spreading of epidemics, information flow and the synchronization of collective behaviour across networks[7,8]. For example, the small-world connectivity pattern proved to be the key to understanding the structure of the World Wide Web[9] and how anatomical and functional areas of the brain communicate with each other[10]. Other structural properties of networks came under the microscope soon after[11–13], such as modularity and the concept of structural motifs, all of which helped scientists to characterize and understand the architecture of living and artificial systems, from subcellular networks to ecosystems and the Internet.

The current generation of network research cross-fertilizes areas that benefit from unprecedented computing power, big data sets and new computational modelling techniques, and thus provides a bridge between the dynamics of individual nodes and the emergent properties of macroscopic networks. But the immediacy and the simplicity of the small-world and preferential-attachment models still underpin our understanding of network topology. Indeed, the relevance of these models to different areas of science laid the foundation of the multidisciplinary field now known as network science.

Integrating knowledge and methodologies from fields as disparate as the social sciences, physics, biology, computer science and applied mathematics was not easy. It took several years to find common ground, agree on definitions and reconcile and appreciate the different approaches that each field had adopted to study networks. This is still a work in progress, presenting all the difficulties and traps inherent in interdisciplinary work. However, in the past 20 years a vibrant network-science community has emerged, with its own prestigious journals, research institutes and conferences attended by thousands of scientists.

By the 20th anniversary of the paper, more than 18,000 papers have cited the model, which is now considered to be one of the benchmark network topologies. Watts and Strogatz closed their paper by saying: "We hope that our work will stimulate further studies of small-world networks." Perhaps no statement has ever been more prophetic. ■

Alessandro Vespignani *is in the Network Science Institute and the Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, Massachusetts 02115, USA.*
e-mail: a.vespignani@northeastern.edu

1. Watts, D. J. & Strogatz, S. H. *Nature* **393**, 440–442 (1998).
2. Erdös, P. & Rényi, A. *Publ. Math.* **6**, 290–297 (1959).
3. Milgram, S. *Psychol. Today* **1**, 61–67 (1967).
4. Guare, J. *Six Degrees of Separation* (Vintage, 1990).
5. Barthélémy, M. & Amaral, L. A. N. *Phys. Rev. Lett.* **82**, 3180–3183 (1999).
6. Barabási, A.-L. & Albert, R. *Science* **286**, 509–512 (1999).
7. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. *Rev. Mod. Phys.* **87**, 925–979 (2015).
8. Arenas, A., Díaz-Guilera, A., Kurths, J., Moreno, Y. & Zhou, C. *Phys. Rep.* **469**, 93–153 (2008).
9. Albert, R., Jeong, H. & Barabási, A.-L. *Nature* **401**, 130–131 (1999).
10. Sporns, O., Chialvo, D. R., Kaiser, M. & Hilgetag, C. C. *Trends Cogn. Sci.* **8**, 418–425 (2004).
11. Newman, M. E. J. *SIAM Rev.* **45**, 167–256 (2003).
12. Porter, M. A., Onnela, J. P. & Mucha, P. J. *Not. Am. Math. Soc.* **56**, 1082–1097 (2009); go.nature.com/2jg9dgq
13. Fortunato, S. *Phys. Rep.* **486**, 75–174 (2010).

This article was published online on 19 June 2018.

STRUCTURAL BIOLOGY

A complex story of receptor signalling

G-protein–coupled receptors activate different G-protein types to trigger divergent signalling pathways. Four structures of receptor–G-protein complexes shed light on this selectivity. SEE ARTICLES P.547, P.553 & P.559 & LETTER P.620

MICHAEL J. CAPPER & DANIEL WACKER

About one-third of all drugs, including opioid painkillers, antihistamines and many antipsychotics, target members of a family of proteins called G-protein-coupled receptors (GPCRs)[1]. This reflects the fact that GPCRs are important in almost all aspects of human physiology, and suggests that many more of them will be promising drug targets for numerous diseases. GPCRs span the cell membrane and convert myriad extracellular signals, including neurotransmitter molecules, hormones, and even light, into a cellular response by activating cellular G proteins and other transducer proteins. Four papers[2–5] in this issue help to unravel the mystery of how GPCRs selectively activate a particular group of G proteins known as $G_{i/o}$, and provide clues that might aid the design of improved GPCR-targeting drugs.

Although more than 800 GPCRs are encoded in the human genome, they couple to only a small number of intracellular signal transducers, including 16 Gα proteins[6]. The latter proteins assemble with Gβ and Gγ proteins to form heterotrimeric G proteins. The G-protein complex disassembles on activation by GPCRs, whereupon the various subunits activate different signalling pathways. For instance, stimulatory Gα proteins (known as G_s) increase cellular levels of cyclic AMP molecules, which regulate various cellular processes. Structures of G_s-bound GPCRs have been reported[7,8] that have begun to elucidate the general activation mechanism of Gα proteins, and of G_s in particular. But much less is known about how GPCRs selectively activate inhibitory Gα proteins, which include G_{i1}, G_{i2}, G_{i3} and G_o, and are collectively known as $G_{i/o}$.

The four papers in this issue report structures of $G_{i/o}$-bound GPCRs obtained using cryo-electron microscopy: Koehl *et al.*[2] (page 547) report the structure of the μ-opioid receptor bound to G_{i1}; Draper-Joyce *et al.*[3] (page 559) describe the adenosine A_1 receptor in complex with G_{i2}; García-Nafría *et al.*[4] (page 620) report the $5HT_{1B}$ receptor bound to G_o; and Kang *et al.*[5] (page 553) reveal the structure of the light receptor rhodopsin in complex with G_{i1}. The G-protein activation cycle involves the binding and release of nucleotides to and from the G proteins, and all of the reported structures capture the receptors bound to the nucleotide-free state of their respective G proteins.

In some respects, the four structures are similar to those of the previously published GPCR–G_s complexes[7,8], probably because G_s- and $G_{i/o}$-containing complexes have the same overall conformation at the stage of the G-protein activation cycle captured by the structures. Nevertheless, the $G_{i/o}$-containing structures reveal striking differences at the receptor–G-protein interface when compared with the G_s-containing structures. For example, there are no interactions between the receptors and the Gβ subunits in the $G_{i/o}$-containing structures.

The four structures uncover several key interactions at the GPCR–$G_{i/o}$ interface mediated by the α5 helix — an α-helix structure in the carboxy terminus of Gα subunits. It is

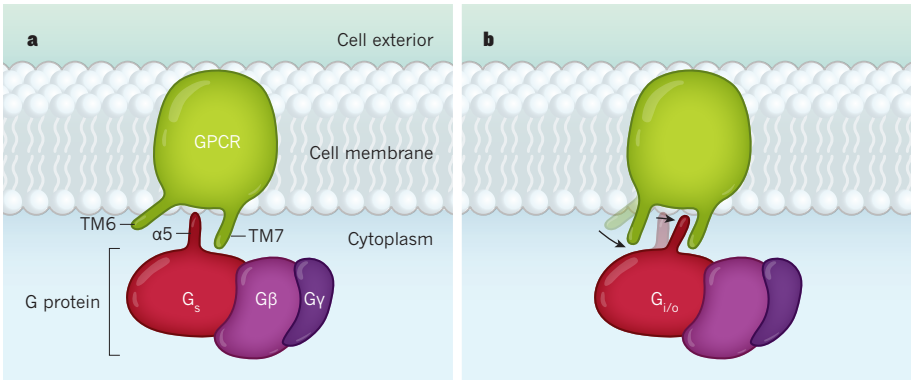


Figure 1 | Structural differences in complexes of G-protein-coupled receptors (GPCRs) with G proteins. GPCRs are transmembrane receptors that activate cellular signalling pathways by binding to G proteins, which have three subunits: α, β and γ. Stimulatory Gα proteins are known as G_s, whereas inhibitory Gα proteins (G_i and G_o proteins) are collectively known as $G_{i/o}$. Many GPCRs selectively bind to G_s or $G_{i/o}$, but the basis of this selectivity was unknown. **a**, This cartoon shows the positions of three α-helices in complexes of GPCRs with G_s-containing G proteins, based on previously reported structures[7,8]. TM6 and TM7 are transmembrane helices in the GPCR, whereas α5 is in the carboxy terminus region of G_s. **b**, Four papers[2–5] now report the structures of GPCRs in complex with $G_{i/o}$ proteins. Compared with **a**, the α5 helices are rotated and moved slightly towards TM7, and away from TM6. The outward displacement of TM6 is smaller than that in **a**. The smaller displacement of TM6 might block the binding of G_s proteins, thus explaining how GPCRs bind selectively to $G_{i/o}$.

known that the binding of this helix to the receptor's cytoplasmic site triggers conformational rearrangements in Gα that cause the release of a nucleotide (GDP) bound to Gα, initiating G-protein activation[9]. The positioning of the $G_{i/o}$ α5 helices in the new structures is different from that of the analogous helices in the GPCR–G_s complexes. Specifically, the $G_{i/o}$ α5 helices are rotated and translated slightly towards transmembrane helix (TM) 7 in the GPCR and away from TM6. Moreover, TM6 is displaced outwards from the receptor core by a smaller amount than occurs in the G_s-bound GPCRs (Fig. 1). The authors of all four papers therefore suggest that the smaller displacement of TM6 might preclude binding of G_s and help to explain how GPCRs can bind selectively to $G_{i/o}$ proteins.

The difference in the positioning of the α5 helices seems to be due to the G_s α5 helices containing bulkier amino-acid residues than those of the $G_{i/o}$ α5 helices. Moreover, Kang *et al.* analysed and compared the amino-acid sequences for TM6 in the G_s- and $G_{i/o}$-coupled receptors, and suggest that the different patterns of hydrophobic and hydrophilic residues observed in the two systems might affect the amount of displacement of TM6, and thus contribute to $G_{i/o}$ specificity.

Comparison of the four GPCR–$G_{i/o}$ structures reveals considerable structural plasticity at the interface. This is not surprising, given that $G_{i/o}$ proteins are engaged by hundreds of GPCRs that have diverse structures and sequences. Draper-Joyce *et al.* thus suggest that G-protein specificity is not necessarily encoded by evolutionarily conserved interactions between specific amino-acid residues, but might be based on "pocket complementarity", in which conformational rearrangements produce regions on the GPCR cytoplasmic site that are conducive to the binding of specific G proteins. Further evidence for this comes from the fact that all the structures of the GPCR–$G_{i/o}$ complexes display markedly smaller GPCR–G protein interfaces than do the structures of GPCR–G_s complexes. This is particularly pronounced for the 5-HT_{1B} receptor–G_o interface surface, which García-Nafría *et al.* report has an area of 822 square ångströms; this compares with 1,260 Å^2 and 1,135 Å^2 for the interfaces in the G_s-bound β_2-adrenergic[7] and adenosine A_{2A} receptors[8], respectively.

Finally, Koehl *et al.* report subtle, yet potentially crucial, differences in the conformations of G_{i1} and G_s that occur during the transition between the GDP-bound and the nucleotide-free states of the proteins. Given that GPCRs catalyse specific structural transitions in specific G-protein subtypes, it is tempting to speculate that the observed conformational differences might also contribute to the G-protein specificity of GPCRs.

"The G-protein specificity is not necessarily encoded by evolutionarily conserved interactions between specific amino-acid residues."

This body of work provides a key step towards delineating the molecular mechanisms by which GPCR conformations drive the activation of one signalling pathway in preference to another. Many more such structures are sure to follow, and will probably reveal structural hallmarks that drive GPCR coupling to other G proteins and signal transducers, such as arrestin proteins. However, as the authors of all four papers point out, these studies provide only snapshots of the G-protein activation pathway, and are thus incomplete. The coupling specificity of GPCRs depends on several factors not addressed by the new structures, including the pre-coupling of G proteins[10] (a preliminary step in which GPCRs and G proteins associate with each other, before actually coupling), and the binding of the GDP-bound form of G proteins[11]. The lifetimes of distinct receptor conformations can also determine the specificity of GPCRs for transducers[12,13], adding a kinetic dimension to GPCR signalling that needs to be considered.

A comprehensive molecular model of GPCR specificity for G proteins and transducers would not only improve our understanding of how GPCRs elicit complicated signals involving multiple, occasionally intersecting, pathways, but also facilitate the design of better drugs that target GPCRs. In particular, it could allow the structure-based design of drugs that selectively activate or inhibit particular signalling pathways, thereby making them safer and more effective than currently available therapeutics.

For example, the painkilling properties of opioid medications such as morphine are thought to arise from the activation of a G_i protein by the μ-opioid receptor, whereas coupling of the receptor to arrestin probably causes the drugs' addictive properties and the — often fatal — depression of respiratory functions. Much effort has thus been dedicated to designing opioid compounds that provide pain relief, but that reduce the risk of addiction or overdose. A flurry of structures of isolated GPCRs has already greatly facilitated the discovery of compounds that bind to the receptors, and that are useful tools for laboratory studies[14]. But it is the structures of GPCR signalling complexes that will allow the rational design of pathway-selective drugs. After all, GPCR signalling is, literally, a complex story. ■

Michael J. Capper *is in the Department of Pharmacological Sciences, and* **Daniel Wacker** *is in the Department of Pharmacological Sciences and the Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.*
e-mail: daniel.wacker@mssm.edu

1. Santos, R. *et al. Nature Rev. Drug Discov.* **16**, 19–13 (2017).
2. Koehl, A. *et al. Nature* **558**, 547–552 (2018).
3. Draper-Joyce, C. J. *et al. Nature* **558**, 559–563 (2018).
4. García-Nafría, J., Nehmé, R., Edwards, P. C. & Tate, C. G. *Nature* **558**, 620–623 (2018).
5. Kang, Y. *et al. Nature* **558**, 553–558 (2018).
6. Milligan, G. & Kostenis, E. *Br. J. Pharmacol.* **147**, S46–S55 (2006).
7. Rasmussen, S. G. F. *et al. Nature* **477**, 549–555 (2011).
8. García-Nafría, J., Lee, Y., Bai, X., Carpenter, B. & Tate, C. G. *eLife* **7**, e35946 (2018).
9. Mahoney, J. P. & Sunahara, R. K. *Curr. Opin. Struct. Biol.* **41**, 247–254 (2016).
10. Andressen, K. W. *et al. FASEB J.* **32**, 1059–1069 (2018).
11. Gregorio, G. G. *et al. Nature* **547**, 68–73 (2017).
12. Wacker, D. *et al. Cell* **168**, 377–389 (2017).
13. Lane, J. R., May, L. T., Parton, R. G., Sexton, P. M. & Christopoulos, A. *Nature Chem. Biol.* **13**, 929–937 (2017).
14. Roth, B. L., Irwin, J. J. & Shoichet, B. K. *Nature Chem. Biol.* **13**, 1143–1151 (2017).

REVIEW

https://doi.org/10.1038/s41586-018-0240-x

Triggers of tree mortality under drought

Brendan Choat[1]*, Timothy J. Brodribb[2], Craig R. Brodersen[3], Remko A. Duursma[1], Rosana López[1,4] & Belinda E. Medlyn[1]

Severe droughts have caused widespread tree mortality across many forest biomes with profound effects on the function of ecosystems and carbon balance. Climate change is expected to intensify regional-scale droughts, focusing attention on the physiological basis of drought-induced tree mortality. Recent work has shown that catastrophic failure of the plant hydraulic system is a principal mechanism involved in extensive crown death and tree mortality during drought, but the multi-dimensional response of trees to desiccation is complex. Here we focus on the current understanding of tree hydraulic performance under drought, the identification of physiological thresholds that precipitate mortality and the mechanisms of recovery after drought. Building on this, we discuss the potential application of hydraulic thresholds to process-based models that predict mortality.

Forests account for approximately 45% of global terrestrial carbon stocks and have a key role in hydrological and nutrient cycles[1,2]. They also provide a wide array of ecosystem services and are vital for maintenance of biodiversity. While forests continue to face pressure from expanding human populations, which drive changes in land use and deforestation, the threat posed by climate change is less easily quantified. Evidence from a range of sources suggests that rising atmospheric CO_2 concentrations have benefited forests, with CO_2 fertilization enabling an increased leaf area index[3], enhanced water-use efficiency[4] and greater uptake of carbon globally[5]. However, extreme climate events, such as heat waves, droughts, fires and storms, have the potential to offset these benefits, causing widespread tree mortality and a net loss of CO_2 into the atmosphere. Although forests are vulnerable to a wide range of extreme climate events, drought and associated disturbances have the greatest effect globally[6]. Recent projections[7] indicate that land surface warming may lead to longer and more intense droughts, which has focused concern on this area of research and the need for accurate predictions of the effects of drought on forest ecosystems. In this Review, we examine the physiological response of trees to drought, focusing on new insights provided by rapid advances in our understanding of the hydraulic function of plants.

Land plants require an efficient long-distance transport pathway to lift water from the soil to the leaves at a rate that satisfies transpiration[8]. In trees, the xylem tissue (wood) supplies water for all aspects of plant function, including photosynthesis, growth and reproduction. Damage to this hydraulic supply network as a consequence of severe water stress has been identified as a key mechanism that is involved in tree mortality during drought[9–11]. Recent experimental work has quantitatively linked hydraulic failure thresholds to plant mortality[12,13], and field studies have demonstrated that hydraulic failure is a primary pathway for extensive canopy death or plant mortality during natural drought events[14–17].

A number of other co-contributing factors may also have a role in the death of trees during natural droughts[18]. In the absence of catastrophic hydraulic failure, partial disruption of water transport and the regulation of water loss from plants during drought may lead to an increased likelihood of mortality through the depletion of carbohydrate reserves used in respiration and increased vulnerability to pests and pathogens[11]. Therefore, even in cases of co-morbidity, plant hydraulic traits occupy a central role in determining survival during drought and the effects of drought on carbon dynamics.

Here, we cover recent progress in our understanding of plant hydraulic response to drought and the physiological mechanisms that govern recovery of hydraulic function after drought. Although recent advances have crystallized our understanding of plant hydraulic function and the consequences of vascular impairment caused by drought stress, many challenges remain. We evaluate recent attempts to integrate the hydraulic traits of plants into process-based models of tree mortality with an emphasis on major knowledge gaps.

Drought and forest mortality

The effect of future droughts will almost certainly be worsened by increases in air temperature associated with global warming; when natural droughts occur they will set in more quickly and be of greater intensity[7]. Higher temperatures will usually result in greater evapotranspiration (the sum of evaporation and plant transpiration), thus drying soil and plants more quickly than would be the case at lower temperatures[19]. Droughts of this nature, termed 'global change-type droughts', have had severe effects on exposed ecosystems including mass tree mortality[20,21].

Globally, drought is the most widespread stress factor that affects forest carbon balance[6] with the potential to cause pronounced depressions in gross primary productivity at regional and continental scales[22,23]. The most notable effects of drought are manifested in regional-scale forest mortality events, which can kill millions of trees within short timescales. Recent high-profile examples include extreme droughts in Texas and California, which are estimated to have killed 300 million and 102 million trees, respectively[24–26]. Mass tree mortality due to drought is not restricted to arid regions, having been documented across many forest biomes including cool temperate and tropical forests[14–16,27,28]. In tropical northern Australia, the sudden die-off of more than 7,000 ha of mangrove forest in 2015 was attributed to drought and extreme temperatures[28]. Although such concentrated mortality events are yet to be observed in many of the world's most productive tropical ecosystems, drought events in tropical rainforests (for example, the 2005 Amazon drought) have resulted in marked increases in stem mortality and loss of aboveground biomass[29]. Mortality is often skewed towards young trees but recent evidence suggests that large, old trees are also vulnerable[30,31]. Loss of large trees is particularly concerning because they have a critical ecological role and have the largest biomass and storage of carbon.

[1]Hawkesbury Institute for the Environment, Western Sydney University, Richmond, New South Wales, Australia. [2]School of Biological Sciences, University of Tasmania, Hobart, Tasmania, Australia. [3]School of Forestry and Environmental Studies, Yale University, New Haven, CT, USA. [4]PIAF, INRA, Université Clermont Auvergne, Clermont-Ferrand, France. *e-mail: b.choat@westernsydney.edu.au

Against this backdrop, it is essential to improve the accuracy with which we can predict the response of trees to drought to understand the resilience of forests under future climate regimes. At present, mortality is not well-represented in vegetation models, owing mostly to gaps in our understanding of physiological mechanisms and a lack of appropriate thresholds with which to parameterize these models. We therefore turn our attention to how these problems may be resolved.

Drought and hydraulic failure in trees

As with all vascular plants, trees prevent desiccation injury by using an intricate plumbing system of hollow dead cells (vessels or tracheids) to transport water from the soil to the leaves. Xylem transport relies on an elegant mechanism whereby liquid water is held under tension, enabling trees to lift vast volumes of water to the canopy at little energetic cost[32]. However, liquid water under tension exists in a metastable state, similar to that of a superheated liquid[33]. In this state, water is prone to cavitation, a sudden phase change from liquid water to gas that creates a bubble (embolism). These gas emboli block water flow through xylem conduits and reduce the delivery of water to the canopy and regenerative tissues (that is, apical and cambial meristems)[8]. Drought leads to higher xylem tensions and an increased probability that emboli will spread throughout the xylem network causing systemic vascular dysfunction[12,34].

Phases of drought stress and the response of plants

During drought, reduced precipitation leads to declines in soil moisture, which are often accompanied by higher temperatures and increased evaporative demand from the atmosphere. These factors combine to induce water stress in plants, which is manifested as increased tension in the xylem sap. Water stress is measurable in plants as xylem water potential (Ψ_x), a variable that is primarily determined by pressure in the xylem fluid and becomes increasingly negative during drought[32,35]. As plants desiccate, the loss of cell turgor causes stomatal pores on the leaf surface to close, markedly slowing plant dehydration and the rate of decrease in Ψ_x. Most recent studies indicate that stomata in trees close before reaching the threshold Ψ_x at which significant cavitation is initiated, despite the negative consequences of stomatal closure[36–38] (Fig. 1). On short time scales, these consequences include a rapid cessation of photosynthetic CO_2 assimilation, loss of canopy evaporative cooling through transpiration and greater probability of photodamage[39,40]. Over longer time scales, low photosynthetic rates associated with drought-induced stomatal closure can lead to depletion of non-structural carbohydrate pools, which interferes with translocation of sugars through the phloem[11,41,42] and the production of chemical defence compounds needed to prevent herbivory and disease[18,43]. The fact that stomatal closure generally occurs before the initiation of cavitation despite these costs suggests that avoidance of xylem cavitation is of paramount importance for the long-term survival of trees.

After stomatal closure, Ψ_x continues to slowly decrease, becoming more negative as water is lost through cuticular conductance, stomatal leakiness[44] and other tissues such as bark[45]. At the same time, hydraulic conductance may decrease throughout the hydraulic pathway of the plant through a number of biophysical and physiological mechanisms, including reversible collapse of leaf veins[46], regulation of aquaporins in cell membranes[47,48] and the formation of cortical lacuna in fine roots[49]. Rates of water loss during this phase are typically in the order of 100–1,000-fold less than when the stomata are fully open[44] and decreases in Ψ_x are further buffered by the release of internally stored water[50]. If drought persists, Ψ_x will ultimately reach a critical threshold at which emboli begin to propagate through the xylem[8,51]. This process occurs throughout the hydraulic pathway including roots, stems and leaves[34,48,49,52,53] (Fig. 2). Because emboli greatly reduce water delivery to the canopy, this hydraulic dysfunction can cause patchy branch death and pronounced reductions in canopy leaf area[54]. During intense droughts, emboli spread throughout the water transport network, causing systemic failure of the vascular system[55]. In the face of continuing drought and high evaporative demand, systemic vascular dysfunction may cause rapid mortality of the whole plant through desiccation[12,15,16] and death of the meristematic tissue in the cambium and apical meristems.

Hydraulic traits of trees and adaptations to drought

The risk of hydraulic failure is an unavoidable consequence of transporting water under tension, and thus forms a fundamental axis of selection in the evolution of trees[56,57]. Strategies to preserve the integrity of the plant vascular system in trees are diverse, but all revolve around a simple framework defined by two constraints: (1) the physical limits of the vascular system; and (2) the capacity to maintain plant water potential within these functional limits. These two attributes dictate how rapidly plant tissues will dehydrate during a drought and the specific thresholds at which water stress will translate into hydraulic failure and mortality[37,58].

Although it is possible to characterize a general sequence of events that describe the response of vascular plants to drought, the traits that define this response vary across species and environments[59–61]. Recent studies have illustrated the enormous variation in vulnerability to xylem cavitation across tree species, with changes in xylem vulnerability correlated to mean annual precipitation and aridity of their growth environment[60]. Species are typically compared by the Ψ_x value at which a 50% loss of hydraulic conductance occurs (Ψ_{50}), although other reference points may have more physiological importance, for example, Ψ_{88} (Fig. 1). Differences in vulnerability are driven by the anatomical features of the xylem, including conduit dimensions, network organization and the porosity of primary cell walls (pit membranes) that limit the spread of gas between conduits[62,63]. These features control the critical Ψ_x at which gas will penetrate pit membranes, causing cavitation in adjacent conduits and the spread of embolism through the xylem[8]. However, vulnerability to cavitation does not determine drought tolerance in itself. The probability of reaching the critical threshold and the length of time it takes for this to occur are determined by the interaction of a number of associated physiological and morphological traits (Fig. 3).

The multi-dimensional nature of such trait interactions has enabled vascular plants to inhabit nearly every terrestrial habitat on Earth and enabled a huge number of possible morphological and physiological solutions to tolerating drought. For instance, variation in the vulnerability of plants is often high within communities, particularly in drier habitats, indicating that vulnerability and aridity are decoupled in some cases[64]. This decoupling results from water-stress avoidance strategies that are used by some species, such as deep root systems or drought deciduousness, that allow them to maintain a higher Ψ_x during drier periods. Although this complexity makes the development of models challenging, a suite of well-studied traits that are mechanistically linked to drought tolerance have now emerged (Supplementary Table 1) and represents a promising direction for future research. Recent analyses have suggested that these traits often vary in a coordinated fashion that allows the benefits of photosynthetic carbon gain to be balanced against the risks of a decrease in Ψ_x and the occurrence of hydraulic failure[59,65]. Thus, much of the complexity of trait interactions may collapse onto a single axis that defines a spectrum of drought tolerance strategies[66].

Ultimately, we are interested in predicting when a plant will die as a result of drought stress. Vulnerability to cavitation has emerged as a key physiological trait that is associated with mortality, and hydraulic failure represents a critical point in the drought response pathway. Species-specific tree hydraulic limitations provide a powerful mechanistic explanation for the observation that drought mortality is occurring across forest biomes, independent of the mean rainfall at any site. A recent data synthesis demonstrated that the majority of plant species converge to narrow hydraulic safety margins, that is, the buffer between minimum water potential experienced by the plant (Ψ_{min}) and the threshold Ψ_x for rapid loss of vascular function caused by cavitation[60]. Because Ψ_{min} integrates many important aspects of plant structure (for example, rooting depth) and physiology (for example, stomatal

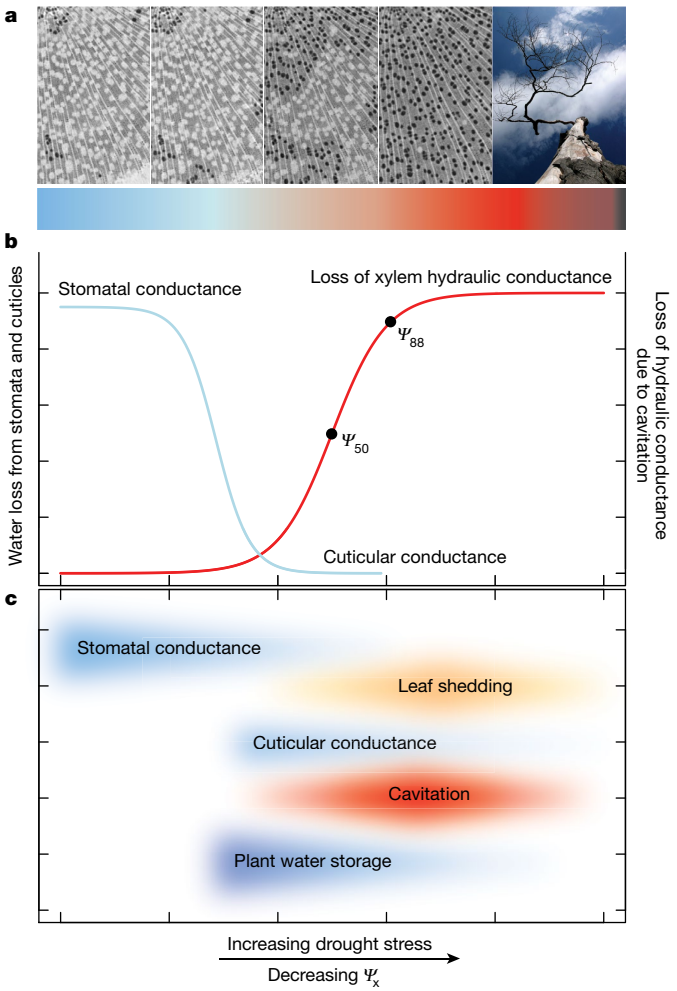


Fig. 1 | Phases of drought response in plants. a, Time series of transverse slices through the xylem tissue obtained by X-ray microtomograpy show the spread of gas emboli through the xylem with increasing drought stress (left to right). In each slice, water-filled vessels are seen as bright circles whereas vessels that contain gas emboli are black. During severe drought, almost all vessels become gas-filled, which leads to whole-plant mortality (right). **b**, During the first phase, stomata close to limit water loss and delay the decrease in xylem water potential (blue line). After stomata close, water continues to be lost at a much lower rate via cuticular conductance. At a critical threshold, cavitation increases rapidly and gas emboli spread throughout the xylem (red line). Increasing levels of embolism are shown as the proportional loss of xylem hydraulic conductance. 'Vulnerability curve' analysis translates the physics of cavitation to a quantification of species susceptibility to cavitation during exposure to water stress. These mortality thresholds have been found to correspond to between 50% (Ψ_{50}) and 88% (Ψ_{88}) loss of hydraulic function in conifers and angiosperms, respectively. **c**, A general scheme for the magnitude and timing of response processes with increasing drought stress.

behaviour) in relation to the environment, the narrow safety margins that are found across forest types offer an important insight into plant ecology, one that suggests that the hydraulic strategies of plants are finely tuned to their environment, allowing for maximum carbon gain but exposing plants to the risk of hydraulic failure during drought. It also suggests a generally 'risky' strategy in which plants have limited physiological potential to respond to rapid changes in the environment. This exacerbates the threat posed by increased occurrences of extreme drought under climate change. Indeed, drought mortality events across forests from a broad geographical and climatic range have been linked quantitatively to hydraulic traits and xylem cavitation[67]. Examples come from tropical rainforest[68], temperate forests[14,16], chaparral[15,17] and desert woodlands[11]. Many of these studies show differential species mortality within each forest type, allowing us insights into the potential winners and losers under future drought regimes.

Plasticity and genetic variation in hydraulic traits

Although much is known about the variability in hydraulic traits among plant species, far fewer data are available to quantify within-species variation. The capacity of trees to alter phenotype (that is, phenotypic plasticity) and the amount of genetic diversity within a population are key variables for the ability of species to cope with rapid climate change. It is unlikely that trees will be capable of adapting to sudden increases in the aridity of their environment through evolutionary mechanisms, because of their long generation cycle and inability to migrate away from stress. On the other hand, adaptive plasticity of hydraulic traits may enable the acclimatization of entire populations within the necessary timescales. Quantifying the extent of plasticity in hydraulic traits is therefore an essential component for the prediction of the tolerance ranges and resilience to drought of different species. However, comprehensive datasets that examine the genetic variation and phenotypic plasticity of vulnerability to cavitation have only recently become available and are limited to a few species. A study of 513 genotypes of the widespread pine species *Pinus pinaster* showed low genetic variation of Ψ_{50} between climatically contrasting populations and very limited phenotypic plasticity[69]. These results suggest that Ψ_{50} may be a canalized trait in pines, with little capacity to enable short-term acclimatization and adaptive plasticity. Angiosperm species have a higher potential for phenotypic plasticity and adaptive variation between populations[70], although the observed shifts in vulnerability are often small in magnitude relative to changes in Ψ_x that are expected to occur during severe drought.

Long-term manipulative experiments suggest that structural acclimatization, that is, changes in the allocation pattern between

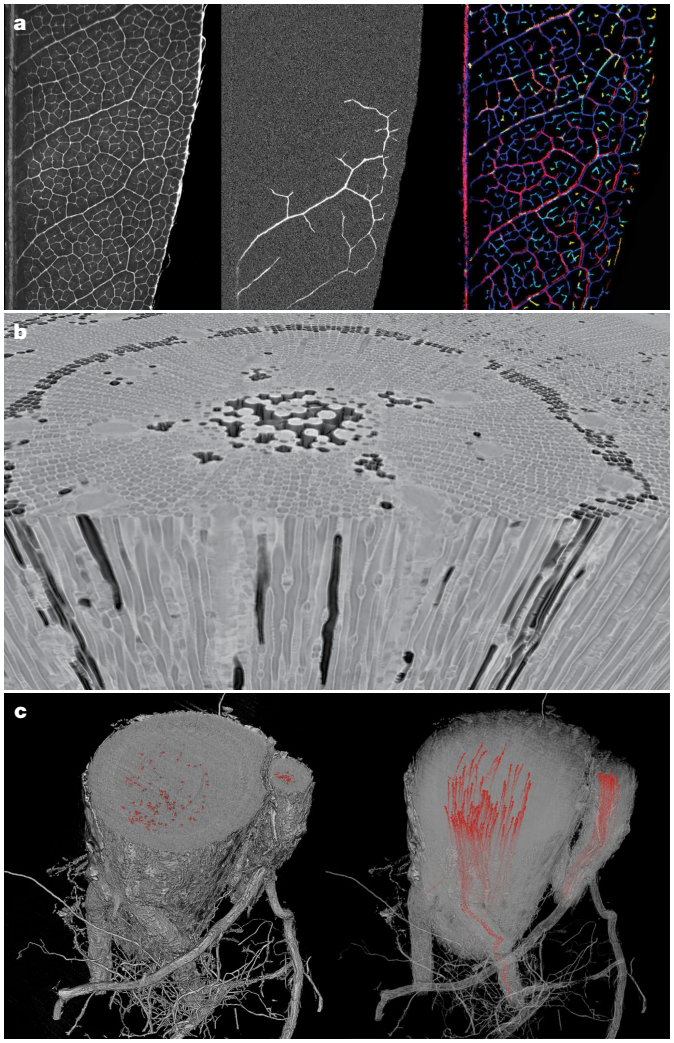


Fig. 2 | Non-invasive imaging techniques have provided new insights into embolism formation and spread in the xylem. a, Mapping the spread of embolism in leaf vein networks during dehydration with transmitted light. Left, transmitted light images highlighting the vein network. Middle, image subtraction reveals embolism propagating from the midrib into the secondary and tertiary venation. Right, a colour map of all cavitation events recorded during desiccation. **b**, Three-dimensional rendering from a X-ray micro-computed tomograph of a pine stem. Embolized tracheids can be seen clearly as a black void space surrounded by water-filled (grey) tracheids. **c**, Part of a root system rendered from a micro-computed tomograph showing embolized xylem vessels (red) in the main root axis and lateral roots during dehydration. Right, the root tissue has been made transparent to illustrate the pathway of embolized vessels. Images in **a** were reproduced with permission from Brodribb et al.[52].

water-absorbing, -conducting and -transpiring tissues, is almost certainly the dominant process by which plants adjust their hydraulic systems in response to drought[71]. Reductions in the leaf to sapwood area ratio and shoot to root ratio result in a greater capacity to supply water to the leaves and limit the drop in Ψ_{min}, consistent with homeostasis of water transport, however these changes come at the cost of reductions in productivity[72]. Reductions in the leaf to sapwood area ratio result in the maintenance of a higher Ψ_x and a greater capacity to supply water to the leaves[73]. These results are consistent with studies of intraspecific variation in hydraulic architecture across aridity gradients, which show changes in morphology and allocation patterns; however, little evidence of adaptive variation in vulnerability to cavitation has been found in these studies, even in species with a wide climate envelope[74,75]. Further studies are clearly required to determine whether these patterns can be generalized, particularly in angiosperms, and what role the plasticity of hydraulic traits may have in the capacity of plants to survive increased aridity.

Predicting mortality from hydraulic thresholds

Predictions of drought-induced forest mortality require a detailed understanding of the physiological underpinnings of tree death. Accordingly, this topic has received much attention in recent years and substantial progress has been made in our understanding of the mechanisms of tree mortality[11,12]. It is clear that drought-associated forest mortality is complex and a number of interdependent mechanisms have important roles in this process. These mechanisms include failure of water transport in the xylem, depletion of carbohydrate reserves over prolonged drought[41,42] and increased vulnerability to pests and pathogens[18,76]. All mechanisms of drought-associated mortality revolve around the effects of stomatal closure and increasing xylem tension during water shortage. Hydraulic failure is the most fully elaborated mechanism and currently holds the most promise for predictive models. It is a relatively well-understood biophysical process that is amenable to modelling[77], with failure thresholds that can be readily established for a given species or population[69,70]. Accuracy and confidence in the vulnerability thresholds that are chosen to represent different species in predictive models are absolutely critical. Recent technical and theoretical advances in the science of plant hydraulics have provided new certainty in the quantitative nature of hydraulic failure[34,55,78]. We thus focus on hydraulic traits as a means to understand and predict patterns of tree mortality in response to drought while emphasizing that other thresholds can be incorporated as our understanding of them improves. Indeed, incorporation of hydraulics thresholds, such as turgor loss and stomatal closure, should assist greatly in predicting carbon dynamics under drought.

Measuring hydraulic thresholds to mortality

As noted above, tree species exhibit a large range in xylem vulnerability[60,61]. Clear links between xylem cavitation and tree death have been established in pot studies[10,12,13] and natural systems[15,79,80], suggesting that xylem vulnerability should undergo strong selection in species exposed to episodic water stress. Evidence of selection, or ecological sorting, can be seen in the distribution of species with regard to strong correlations between aridity and xylem vulnerability[10,56,64]. The exciting implication of this work is that the vulnerability of xylem tissue to cavitation provides a measurable index of the capacity of a species to tolerate water stress during drought[77].

Our ability to predict the level of water stress at which a plant will die based on functional traits has advanced considerably in recent years. The concept of a 'lethal water potential' for a given plant species or population has existed for some time, but only recently has hydraulic vulnerability been quantitatively linked to mortality[12]. In conifers, the Ψ_{50} of the stem xylem is strongly related to their 'minimum recoverable water potential', essentially a physiological point of no return[12,81]. By contrast, the lethal water potential of angiosperm species is correlated with more complete hydraulic dysfunction, representing 80–100% loss of xylem hydraulic conductance (Ψ_{88})[10,13]. The disparity in mortality thresholds between conifers and angiosperms may be related to the fundamental difference in xylem structure between these two groups and goes some way to explaining the generally larger hydraulic safety margins and more conservative stomatal behaviour exhibited by conifer species[60,82]. Other hydraulic thresholds for mortality have been proposed, including a sustained loss of hydraulic conductance greater than 60%[80,83]. We note that these studies were based on modelled thresholds of whole-plant conductance in two conifer species and compare well to the Ψ_{50} threshold that has been established for conifers experimentally. A recent data synthesis found that all studies reported a 60% or greater loss in hydraulic conductance at death with a mean loss of 83%[9]. Therefore, although there are some discrepancies between proposed thresholds, it is clear that high levels of xylem embolism are linked with mortality. Modern techniques for the measurement of water potential and non-invasive visualization of embolism are providing more

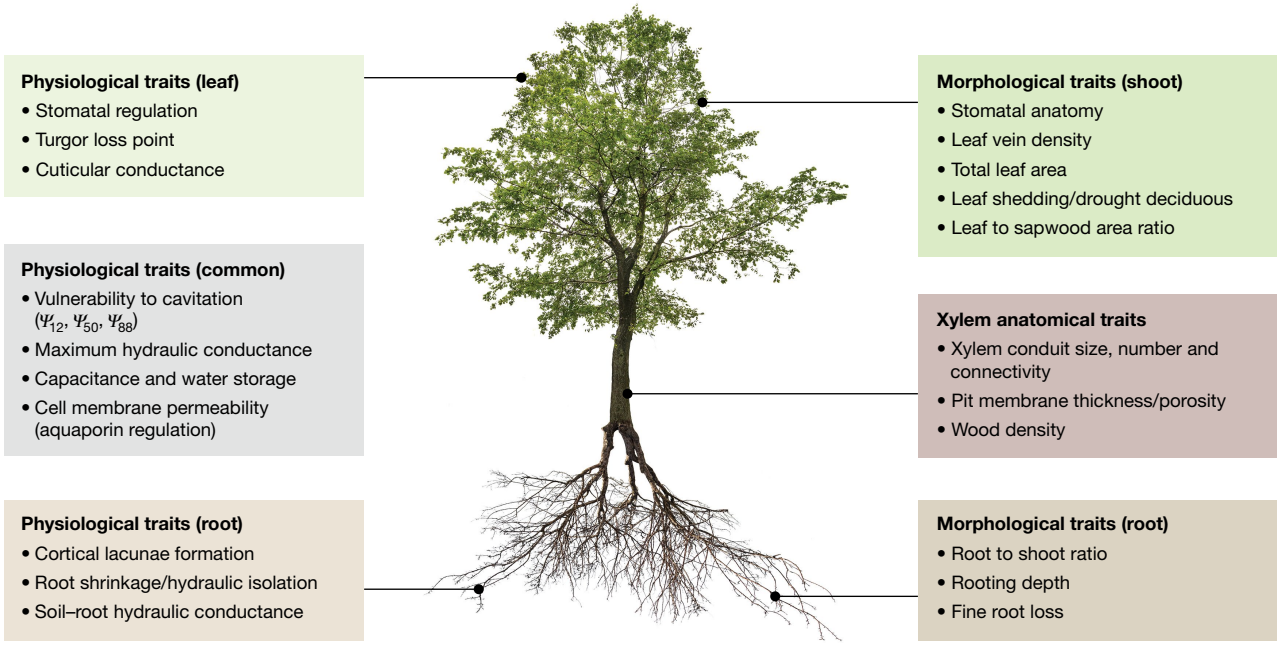


Fig. 3 | Tree hydraulic traits associated with drought-induced mortality. Trees use a variety of interdependent and coordinated morphological, anatomical and physiological traits to mitigate water loss and the development of increasingly negative xylem sap pressures during drought. This includes tissue-specific traits that function in the unique microenvironment of roots, stems and leaves, as well as traits that are common among most tissue types in trees. Many structure–function relationships exist between traits, for example, variation in xylem anatomical traits (pit membrane porosity, conduit size and connectivity) determine species and population-level vulnerability to cavitation. Note that this figure does not represent an exhaustive list of hydraulic traits relevant to the response of trees to drought and drought-induced mortality.

accurate measures of hydraulic failure thresholds[34,55], which can then be used to parameterize models of tree mortality.

Trait-based models of tree mortality

Although the vulnerability of xylem to cavitation defines a threshold in water stress beyond which tree mortality will occur, the key issue for predicting mortality is the ability to translate meteorological data (for example, precipitation or evaporative demand) into plant water content or xylem tension[58]. This calculation presents a number of challenges: it requires knowledge of the volume of water that is available to a plant in the soil and internal reservoirs, as well as the rate of water loss through transpiration. Calculating the amount of water in the soil that is available to the plant is made difficult by a paucity of data relating to rooting depth of trees and the architecture of the roots. Simulating the rate of water loss is complicated by the active regulation of transpiration by stomata, differences in cuticular transpiration after stomatal closure and the degree of leaf shedding during drought. To date, most attempts to model or predict mortality use empirical relationships between observed mortality and climate extremes[84,85]. Such empirical relationships, although they provide insights into the current drivers of mortality, may not function well in the future if plant sensitivities change over time or if novel climate conditions occur. For example, rising CO_2 concentrations may alleviate drought stress, whereas rising temperatures may exacerbate it[86]. Empirical relationships may also fail in regions where long-term shifts towards a novel, drier climate are occurring. Process-based models are thus highly desirable[87].

Recent progress in the understanding of the hydraulic mechanisms that lead to mortality, and quantification of the key plant traits that are involved, has led to the incorporation of plant hydraulics in a range of process-based vegetation models[83,88–90]. The key elements of such models are a description of the soil-to-leaf hydraulic pathway, incorporating soil, root, xylem and stomatal conductances, whole-plant capacitance and the vulnerability of the xylem to cavitation. Important plant traits that are required for parameterization include the response of stomata to decreasing water potential, the point at which leaf turgor is lost, saturated xylem hydraulic conductance and water potential thresholds of vulnerability to cavitation. Recent compilation efforts have made data for these traits available for a wide range of species[60,91] and enabled hydraulic traits to be related to other aspects of the plant economic spectrum[88,89]. Incorporating trait variation in drought sensitivity among species or genotypes, and relating this variation to plant water-use strategies and other plant properties, promises to be an effective way forwards.

Nonetheless, critical gaps remain in our ability to describe the hydraulic pathway and its eventual failure. Here we draw attention to several gaps that hinder model development and parameterization and that have received comparatively little attention: (1) the dynamics of canopy leaf area during drought; (2) the dependence of plant water status on soil water potential; and (3) the process of plant desiccation in very dry soil, when root water uptake is no longer possible.

First, leaf shedding occurs in many ecosystems during drought[92], and can mitigate water stress to the remaining foliage[93], slowing the rate of desiccation (the 'hydraulic fuse' hypothesis[53]). However, drought deciduousness is, as yet, poorly captured in models[94]. Recently, it has been demonstrated that representing vegetation as a set of competing plant types with varying degrees of drought deciduousness leads to a marked improvement in modelled leaf area dynamics in Central America; this hydraulics-based approach holds considerable promise for model improvement[89].

Second, a key component of the hydraulic pathway is the relationship between plant water status (represented by pre-dawn plant water potential, Ψ_{pd}) and soil water availability (represented by soil water potential, Ψ_{soil}). Simple models that treat soil water as a single bucket generally fail to capture this relationship; it appears that models need to incorporate vertical gradients in soil moisture potential, the distribution of roots and changing soil–root resistance with soil drying[88,95]. However, it is also commonly observed that co-occurring species can have different Ψ_{pd} when Ψ_{soil} is the same[96], and this difference cannot always be explained by rooting distributions. The Ψ_{pd} can be lower than Ψ_{soil} if overnight equilibration is insufficient[89], or considerable amounts of

night-time transpiration occur[97]. On the other hand, root shrinkage or cortical deformation in dry soil can create an air gap around roots[49,98], preventing equilibration and leading to a Ψ_{pd} that is less negative than the Ψ_{soil}. Foliar water uptake can also lead to shoot rehydration disequilibria with Ψ_{soil} and these effects can be important for arid environments[99]. There has been little focus to date on the ability of models to replicate measurements of Ψ_{pd}.

Third, it is clear that processes that occur after stomatal closure are important, however these processes have been given less consideration in models[58,100]. Stomata generally close well before the thresholds for pronounced cavitation are reached[37,59,101]. Further increases in cavitation will occur only if models also represent plant water loss when stomata are closed. In very dry soil, the uptake of water by roots is no longer possible and plants have to rely on their internal water storage for survival[16,58]. Plant water storage (an absolute amount of water) is distinct from the capacitance (the slope of water content versus potential). Whereas capacitance is often incorporated into models because it determines the dynamics of the water potential in leaves[102], plant water storage is not. The depletion rate of the plant water store when stomata are (nearly) closed is governed by the plant leaf area (discussed above), cuticular conductance[44] and, probably less importantly, water loss through the bark[45]. Comparatively few measurements are available on cuticular conductance and plant water storage[58], hindering model parameterization for this final phase towards drought mortality. At this point, very few process-based models that have aimed to simulate drought mortality incorporate both plant water storage and cuticular conductance. A notable exception is provided by the recently developed SurEau model[37].

In summary, trait-based models that incorporate our best understanding of hydraulic processes hold promise for predicting plant mortality in response to drought. We have highlighted three processes that deserve further attention, both in terms of model development and compilation of necessary data for parameterization. There are likely to be other problems as well, perhaps even larger ones—it is worth noting that no purely process-based model has yet been successful in predicting tree mortality[90]. Model testing must thus continue before we can defensibly use these models for forecasting. We recommend a close coordination between experimentalists and modellers to improve models and their evidence base.

Recovery of hydraulic capacity

Although much attention has been devoted to determining the physiological basis of tree mortality during drought, it is equally important to understand the processes of recovery in trees that survive drought. What are the effects of drought on the hydraulic function of plants and how quickly can plants recover to the pre-drought levels of physiological performance? Predicting the resilience and recovery of forests is complicated by the predisposition of trees that enter a drought event, which includes the cumulative effects of previous water deficit, pest outbreaks and forest demographics, as well as potential delays between stress events that influence survival, mortality and recovery processes[16,58,103,104]. Recovery of trees after drought is therefore complex, dynamic and determined by at least (1) the degree of damage to the apical and cambial meristematic tissues; (2) the functional status of the remaining hydraulic pathway; (3) the overall health of trees (that is, the remaining foliage and roots); and (4) the water, non-structural carbohydrates and nutrients that are available during the recovery phase.

Mild drought stress does not typically result in high levels of cavitation, although it may result in the transient and easily reversible loss of hydraulic capacity that is associated with mechanisms such as conduit collapse in leaf veins and aquaporin regulation of cell membrane permeability[46–48]. In cases in which drought stress has caused considerable hydraulic dysfunction without mortality, hydraulic recovery could occur by two mechanisms: new wood formation or by refilling embolized conduits. Regrowth of the xylem appears to be the primary means by which trees recover hydraulic capacity after drought[12,81]. This straightforward process involves the addition of new conduits (vessels or tracheids) to outer regions of the xylem through the activity of the vascular cambium. This replaces the hydraulic conductance that is lost by embolized conduits, which may then become permanently occluded by gums or tyloses[105]. In cases in which drought has caused considerable death of aboveground biomass, recovery may be facilitated by resprouting of stems from epicormic buds or lignotubers[106]. The prevalence of resprouting is highly variable among tropical and temperate forest species and is ultimately dependent on the protection and survival of the meristematic tissue to produce new shoots[106,107].

An alternative mechanism for recovery of hydraulic capacity, which would be far more rapid than regrowth, is refilling of embolized xylem conduits. Much research has focused on the potential active mechanisms by which plants could refill embolized xylem conduits after drought[108–110]. Springtime refilling following freeze–thaw cycles that produce embolism over winter is well-documented and apparently dependent on the positive pressure that is generated in the roots or stems of deciduous angiosperms[111–113]. A number of studies have shown rapid refilling of embolized vessels after mild drought in herbaceous species[114,115] and previous studies have suggested that daily cycles of cavitation and refilling are common in some tree species[116,117]. However, recent work has cast doubt on this phenomenon and convincing evidence for short-term refilling after drought in large trees is generally lacking[108]. It appears unlikely that trees can establish the positive pressure that is required to remove emboli in transpiring tissues many metres above the soil surface, making refilling under sustained tension within the current theoretical framework thermodynamically untenable[109]. Although it is thought that some woody species can remove drought-induced emboli within hours or days after a soil-saturating event in combination with non-transpiring conditions, this has only been documented unambiguously in grapevines, which are well-known for their capacity to produce considerable root pressure[118,119]. Other studies using non-invasive imaging have failed to provide evidence of refilling in woody species after drought[51,120,121]. Thus, we believe it is unlikely that refilling in trees is a common mechanism for rapid recovery of hydraulic capacity after drought, although further experimental studies are required to confirm this.

The rate of hydraulic recovery after exposure to drought is largely is dictated by the extreme of negative water potential that is reached, and the amount of time that is spent at this extreme. Drought recovery after rainfall occurs on very short time scales if high levels of cavitation have not taken place[122] with the rapid opening of stomata to fix new carbon from the atmosphere as plants re-hydrate. However, in cases in which cavitation thresholds are breached, the recovery of photosynthesis is much slower[122] and proceeds in coordination with the onset of the restoration of hydraulic condutance from the soil to the canopy[12,81]. Growth of new xylem to replace compromised tissues requires long-distance signalling from the roots and leaves[123], and xylogenesis is affected by the post-drought environment because of sensitivity to temperature, plant growth regulators, carbohydrate pools and water availability[124,125]. Following drought, trees invest substantial carbon resources into rebalancing the root to shoot ratio[126], increasing fine root biomass, and exploring deeper regions of the soil profile to recover plant water status[127]. Because meristematic tissue is dependent on adequate water availability and phloem transport to establish the turgor that is necessary for xylem cell expansion[128], xylem development during drought conditions is often markedly reduced. Thus, functional water transport and signalling pathways must be in place to coordinate these events and initiate recovery.

The way forwards

Hydraulic physiology is central to our understanding of how trees respond to drought, and the pathways leading to drought-induced mortality. Because the hydraulic system is fundamentally linked to carbon balance through stomatal regulation, a mechanistic understanding of plant hydraulic function should greatly improve modelling of vegetation dynamics under water-limiting conditions. With recent progress in methodology, we now stand at an exciting threshold in the field.

There is potential for great leaps in our understanding of plant hydraulic function and our ability to quantitatively link these physiological mechanisms to forest ecology. Despite this, many important challenges remain. Below, we outline research priorities and key knowledge gaps that constrain our ability to predict drought-induced mortality and recovery after drought.

In terms of manipulative drought experiments, it is clear that studies incorporating long-term droughts on large trees are essential. Thresholds for hydraulic failure have typically been determined on small plants with intense drought treatments and important differences may exist in large trees exposed to droughts of greater duration. This is particularly applicable to determining the time a tree must spend at or below a particular threshold in order for mortality to occur. Long-term experiments will also provide an improved understanding of how tree water relations and carbon balance interact to cause mortality during prolonged droughts.

Better monitoring of plant water potential in communities under drought is also essential in this context. The importance of Ψ_{min} as a parameter cannot be understated, since the Ψ_x reached by plants is what largely determines the probability that hydraulic failure occurs. It integrates many aspects of plant structure and physiology, and their interaction with climatic and edaphic variables[78]. Other thresholds, such as the point at which leaf turgor is lost and stomatal closure occurs, can be predicted for a species based on continuous Ψ_x data, although seasonal acclimatization in these parameters needs to be taken into account. Unfortunately, datasets of Ψ_{min} are typically patchy because of the laborious onsite measurement techniques that are involved. Continuous remote monitoring of Ψ_x is now becoming possible because of the development of a new generation of wireless sensors that can be deployed at remote sites[129,130]. This will enable a much higher resolution of Ψ_{min}, the true maximum level of stress that trees are exposed to during natural droughts, and how long they spend at a given Ψ_x.

Improved methodology for the measurement of vulnerability to cavitation is also being developed to enable in situ measurement of hydraulic thresholds[52,55] and more rapid phenotyping of thresholds across species and populations[69]. Non-invasive imaging techniques, such as X-ray micro-computed tomography and magnetic resonance imaging, are providing new insights into plant hydraulic function and response to drought. Although these techniques are often costly or difficult to access, they provide unparalleled spatial and temporal resolution for observations in living, intact plants. Further work in this area will be vital to unravel some persistent mysteries of plant vascular transport including the phenomenon of embolism repair after drought and the degree to which it is active in woody plants. Improvements in techniques are also essential for the quantification of plasticity within species and their capacity to acclimatize or adapt to drier conditions. A recently developed optical technique shows great promise to facilitate cost-effective high-throughput measurements of vulnerability to cavitation in leaves, stems and roots[34]. Methods that use centrifugal force to generate vulnerability curves also enable rapid phenotyping[69] provided the appropriate methodological precautions are observed[131].

The use of whole-tree techniques, for example, sap flow, is necessary to provide datasets that interface with databases of tissue-level traits. Measuring sap flow and consolidating existing sap flow datasets worldwide (for example, the Sapfluxnet project)[132] represent important steps in linking tissue-level hydraulic traits to water fluxes and testing large-scale model predictions of the effect of water stress on plants and ecosystems. At a broader scale, remote-sensing tools will be essential to monitor tree mortality and the dynamics of drought recovery. Although there are issues to be solved with detection of mortality using satellite-based sensors[133], aircraft-based sensors are delivering better resolution of water stress and tree death at stand and regional scales[24]. Although the goal of attaining accurate model predictions of tree mortality due to drought remains elusive, resolution of the challenges outlined above represents a clear path forward. Success in this area will require direct collaboration between experimentalists and modellers, as the effective parameterization of process-based models depends on the acquisition and sharing of often hard-won data (for example, rooting depth, cuticular conductance and cavitation resistance). Finally, we emphasize that hydraulic failure is not the only pathway to mortality associated with drought but rather the most tractable to address with process-based models at this time. The future integration of other physiological thresholds, as well as interactions with pests and pathogens, is the natural course along which we should proceed.

Received: 19 October 2016; Accepted: 2 May 2018;
Published online 27 June 2018.

1. Bonan, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science* **320**, 1444–1449 (2008).
2. Pan, Y. et al. A large and persistent carbon sink in the world's forests. *Science* **333**, 988–993 (2011).
3. Zhu, Z. et al. Greening of the Earth and its drivers. *Nat. Clim. Change* **6**, 791–795 (2016).
4. Keenan, T. F. et al. Increase in forest water-use efficiency as atmospheric carbon dioxide concentrations rise. *Nature* **499**, 324–327 (2013).
5. Wenzel, S., Cox, P. M., Eyring, V. & Friedlingstein, P. Projected land photosynthesis constrained by changes in the seasonal cycle of atmospheric CO_2. *Nature* **538**, 499–501 (2016).
6. Reichstein, M. et al. Climate extremes and the carbon cycle. *Nature* **500**, 287–295 (2013).
7. Trenberth, K. E. et al. Global warming and changes in drought. *Nat. Clim. Change* **4**, 17–22 (2014).
8. Tyree, M. T. & Zimmermann, M. H. *Xylem Structure and the Ascent of Sap* (Springer, New York, 2002).
9. Adams, H. D. et al. A multi-species synthesis of physiological mechanisms in drought-induced tree mortality. *Nat. Ecol. Evol.* **1**, 1285–1291 (2017).
10. Kursar, T. A. et al. Tolerance to low leaf water status of tropical tree seedlings is related to drought performance and distribution. *Funct. Ecol.* **23**, 93–102 (2009).
11. McDowell, N. et al. Mechanisms of plant survival and mortality during drought: why do some plants survive while others succumb to drought? *New Phytol.* **178**, 719–739 (2008).
This study introduces a theoretical framework for understanding physiological mechanisms that underpin drought-induced mortality in trees.
12. Brodribb, T. J. & Cochard, H. Hydraulic failure defines the recovery and point of death in water-stressed conifers. *Plant Physiol.* **149**, 575–584 (2009).
This study quantitatively links hydraulic failure thresholds to whole-plant mortality.
13. Urli, M. et al. Xylem embolism threshold for catastrophic hydraulic failure in angiosperm trees. *Tree Physiol.* **33**, 672–683 (2013).
14. Nardini, A., Battistuzzo, M. & Savi, T. Shoot desiccation and hydraulic failure in temperate woody angiosperms during an extreme summer drought. *New Phytol.* **200**, 322–329 (2013).
15. Venturas, M. D. et al. Chaparral shrub hydraulic traits, size, and life history types relate to species mortality during California's historic drought of 2014. *PLoS ONE* **11**, e0159145 (2016).
16. Anderegg, W. R. et al. The roles of hydraulic and carbon stress in a widespread climate-induced forest die-off. *Proc. Natl Acad. Sci. USA* **109**, 233–237 (2012).
17. Davis, S. D. et al. Shoot dieback during prolonged drought in *Ceanothus* (Rhamnaceae) chaparral of California: a possible case of hydraulic failure. *Am. J. Bot.* **89**, 820–828 (2002).
18. McDowell, N. G. et al. The interdependence of mechanisms underlying climate-driven vegetation mortality. *Trends Ecol. Evol.* **26**, 523–532 (2011).
19. Duan, H. et al. Elevated $[CO_2]$ does not ameliorate the negative effects of elevated temperature on drought-induced mortality in *Eucalyptus radiata* seedlings. *Plant Cell Environ.* **37**, 1598–1613 (2014).
20. Allen, C. D., Breshears, D. D. & McDowell, N. G. On underestimation of global vulnerability to tree mortality and forest die-off from hotter drought in the Anthropocene. *Ecosphere* **6**, 129 (2015).
21. Carnicer, J. et al. Widespread crown condition decline, food web disruption, and amplified tree mortality with increased climate change-type drought. *Proc. Natl Acad. Sci. USA* **108**, 1474–1478 (2011).
22. Ciais, P. et al. Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature* **437**, 529–533 (2005).
23. Lewis, S. L., Brando, P. M., Phillips, O. L., van der Heijden, G. M. F. & Nepstad, D. The 2010 Amazon drought. *Science* **331**, 554 (2011).
24. Asner, G. P. et al. Progressive forest canopy water loss during the 2012–2015 California drought. *Proc. Natl Acad. Sci. USA* **113**, E249–E255 (2016).
25. Moore, G. W. et al. Tree mortality from an exceptional drought spanning mesic to semiarid ecoregions. *Ecol. Appl.* **26**, 602–611 (2016).
26. USDA Forest Service Pacific Southwest Region. *Aerial Detection Surveys Report: Summary for May 15–19* Report No. fseprd506698 (USDA Forest Service, 2016).
27. Allen, C. D. et al. A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *For. Ecol. Manage.* **259**, 660–684 (2010).
This study summarizes forest mortality events associated with drought and heat over the last four decades.
28. Duke, N. C. et al. Large-scale dieback of mangroves in Australia's Gulf of Carpentaria: a severe ecosystem response, coincidental with an unusually extreme weather event. *Mar. Freshw. Res.* **68**, 1816–1829 (2017).

29. Phillips, O. L. et al. Drought sensitivity of the Amazon rainforest. *Science* **323**, 1344–1347 (2009).
30. da Costa, A. C. L. et al. Effect of 7 yr of experimental drought on vegetation dynamics and biomass storage of an eastern Amazonian rainforest. *New Phytol.* **187**, 579–591 (2010).
31. Lindenmayer, D. B. & Laurance, W. F. The ecology, distribution, conservation and management of large old trees. *Biol. Rev. Camb. Philos. Soc.* **92**, 1434–1458 (2016).
32. Slatyer, R. O. *Plant–Water relationships* (Academic, New York, 1967).
33. Debenedetti, P. G. *Metastable Liquids: Concepts and Principles* (Princeton Univ. Press, Princeton, 1996).
34. Rodriguez-Dominguez, C. M., Carins Murphy, M. R., Lucani, C. & Brodribb, T. J. Mapping xylem failure in disparate organs of whole plants reveals extreme resistance in olive roots. *New Phytol.* **218**, 1025–1035 (2018).
35. Scholander, P. F., Hammel, H. T., Bradstreet, E. D. & Hemmingsen, E. A. Sap pressure in vascular plants. *Science* **148**, 339–346 (1965).
36. Hochberg, U. et al. Stomatal closure, basal leaf embolism and shedding protect the hydraulic integrity of grape stems. *Plant Physiol.* (2017).
37. Martin-StPaul, N., Delzon, S. & Cochard, H. Plant resistance to drought depends on timely stomatal closure. *Ecol. Lett.* **20**, 1437–1447 (2017).
The absolute limit at which stomata must close to avoid mortality under drought is described.
38. Li, X. et al. Tree hydraulic traits are coordinated and strongly linked to climate-of-origin across a rainfall gradient. *Plant Cell Environ.* **41**, 646–660 (2018).
39. Leigh, A., Sevanto, S., Close, J. D. & Nicotra, A. B. The influence of leaf size and shape on leaf thermal dynamics: does theory hold up under natural conditions? *Plant Cell Environ.* **40**, 237–248 (2017).
40. Powles, S. B. Photoinhibition of photosynthesis induced by visible light. *Annu. Rev. Plant Physiol.* **35**, 15–44 (1984).
41. Mitchell, P. J. et al. Drought response strategies define the relative contributions of hydraulic dysfunction and carbohydrate depletion during tree mortality. *New Phytol.* **197**, 862–872 (2013).
42. Sevanto, S., McDowell, N. G., Dickman, L. T., Pangle, R. & Pockman, W. T. How do trees die? A test of the hydraulic failure and carbon starvation hypotheses. *Plant Cell Environ.* **37**, 153–161 (2014).
43. Dietze, M. C. & Matthes, J. H. A general ecophysiological framework for modelling the impact of pests and pathogens on forest ecosystems. *Ecol. Lett.* **17**, 1418–1426 (2014).
44. Kerstiens, G. Cuticular water permeability and its physiological significance. *J. Exp. Bot.* **47**, 1813–1832 (1996).
45. Oren, R. & Pataki, D. E. Transpiration in response to variation in microclimate and soil moisture in southeastern deciduous forests. *Oecologia* **127**, 549–559 (2001).
46. Zhang, Y.-J., Rockwell, F. E., Graham, A. C., Alexander, T. & Holbrook, N. M. Reversible leaf xylem collapse: a potential "circuit breaker" against cavitation. *Plant Physiol.* **172**, 2261–2274 (2016).
47. McElrone, A. J. et al. Aquaporin-mediated changes in hydraulic conductivity of deep tree roots accessed via caves. *Plant Cell Environ.* **30**, 1411–1421 (2007).
48. Sack, L. & Holbrook, N. M. Leaf hydraulics. *Annu. Rev. Plant Biol.* **57**, 361–381 (2006).
49. Cuneo, I. F., Knipfer, T., Brodersen, C. R. & McElrone, A. J. Mechanical failure of fine root cortical cells initiates plant hydraulic decline during drought. *Plant Physiol.* **172**, 1669–1678 (2016).
50. Borchert, R. & Pockman, W. T. Water storage capacitance and xylem tension in isolated branches of temperate and tropical trees. *Tree Physiol.* **25**, 457–466 (2005).
51. Choat, B., Brodersen, C. R. & McElrone, A. J. Synchrotron X-ray microtomography of xylem embolism in *Sequoia sempervirens* saplings during cycles of drought and recovery. *New Phytol.* **205**, 1095–1105 (2015).
52. Brodribb, T. J. et al. Visual quantification of embolism reveals leaf vulnerability to hydraulic failure. *New Phytol.* **209**, 1403–1409 (2016).
53. Tyree, M. T., Cochard, H., Cruiziat, P., Sinclair, B. & Ameglio, T. Drought-induced leaf shedding in walnut: evidence for vulnerability segmentation. *Plant Cell Environ.* **16**, 879–882 (1993).
54. Rood, S. B., Patiño, S., Coombs, K. & Tyree, M. T. Branch sacrifice: cavitation-associated drought adaptation of riparian cottonwoods. *Trees* **14**, 248–257 (2000).
55. Choat, B. et al. Noninvasive measurement of vulnerability to drought-induced embolism by X-ray microtomography. *Plant Physiol.* **170**, 273–282 (2016).
56. Larter, M. et al. Aridity drove the evolution of extreme embolism resistance and the radiation of conifer genus *Callitris*. *New Phytol.* **215**, 97–112 (2017).
57. Pittermann, J. The evolution of water transport in plants: an integrated approach. *Geobiology* **8**, 112–139 (2010).
58. Blackman, C. J. et al. Toward an index of desiccation time to tree mortality under drought. *Plant Cell Environ.* **39**, 2342–2345 (2016).
A process-based approach is used to model desiccation time to mortality in trees under drought.
59. Bartlett, M. K., Klein, T., Jansen, S., Choat, B. & Sack, L. The correlations and sequence of plant stomatal, hydraulic, and wilting responses to drought. *Proc. Natl Acad. Sci. USA* **113**, 13098–13103 (2016).
60. Choat, B. et al. Global convergence in the vulnerability of forests to drought. *Nature* **491**, 752–755 (2012).
Global synthesis demonstrating a convergence in tree hydraulic safety margins across forest biomes.
61. Maherali, H., Pockman, W. T. & Jackson, R. B. Adaptive variation in the vulnerability of woody plants to xylem cavitation. *Ecology* **85**, 2184–2199 (2004).
62. Lens, F. et al. Testing hypotheses that link wood anatomy to cavitation resistance and hydraulic conductivity in the genus *Acer*. *New Phytol.* **190**, 709–723 (2011).
63. Pittermann, J. et al. The relationships between xylem safety and hydraulic efficiency in the Cupressaceae: the evolution of pit membrane form and function. *Plant Physiol.* **153**, 1919–1931 (2010).
64. Blackman, C. J., Brodribb, T. J. & Jordan, G. J. Leaf hydraulic vulnerability influences species' bioclimatic limits in a diverse group of woody angiosperms. *Oecologia* **168**, 1–10 (2012).
65. Mencuccini, M., Minunno, F., Salmon, Y., Martínez-Vilalta, J. & Hölttä, T. Coordination of physiological traits involved in drought-induced mortality of woody plants. *New Phytol.* **208**, 396–409 (2015).
66. Reich, P. B. The world-wide 'fast–slow' plant economics spectrum: a traits manifesto. *J. Ecol.* **102**, 275–301 (2014).
67. Anderegg, W. R. et al. Meta-analysis reveals that hydraulic traits explain cross-species patterns of drought-induced tree mortality across the globe. *Proc. Natl Acad. Sci. USA* **113**, 5024–5029 (2016).
68. Rowland, L. et al. Death from drought in tropical forests is triggered by hydraulics not carbon starvation. *Nature* **528**, 119–122 (2015).
69. Lamy, J.-B. et al. Limited genetic variability and phenotypic plasticity detected for cavitation resistance in a Mediterranean pine. *New Phytol.* **201**, 874–886 (2014).
70. Schuldt, B. et al. How adaptable is the hydraulic system of European beech in the face of climate change-related precipitation reduction? *New Phytol.* **210**, 443–458 (2016).
71. Mencuccini, M. & Grace, J. Climate influences the leaf area/sapwood area ratio in Scots pine. *Tree Physiol.* **15**, 1–10 (1995).
72. Magnani, F., Mencuccini, M. & Grace, J. Age-related decline in stand productivity: the role of structural acclimation under hydraulic constraints. *Plant Cell Environ.* **23**, 251–263 (2000).
73. Maherali, H. & DeLucia, E. H. Xylem conductivity and vulnerability to cavitation of ponderosa pine growing in contrasting climates. *Tree Physiol.* **20**, 859–867 (2000).
74. Martínez-Vilalta, J. et al. Hydraulic adjustment of Scots pine across Europe. *New Phytol.* **184**, 353–364 (2009).
Comprehensive study of intra-specific variation in hydraulic traits across a broad climatic gradient.
75. Wortemann, R. et al. Genotypic variability and phenotypic plasticity of cavitation resistance in *Fagus sylvatica* L. across Europe. *Tree Physiol.* **31**, 1175–1182 (2011).
76. Hogg, E. H., Brandt, J. P. & Kochtubajda, B. Growth and dieback of aspen forests in northwestern Alberta, Canada, in relation to climate and insects. *Can. J. For. Res.* **32**, 823–832 (2002).
77. Sperry, J. S. & Love, D. M. What plant hydraulics can tell us about responses to climate-change droughts. *New Phytol.* **207**, 14–27 (2015).
78. Delzon, S. & Cochard, H. Recent advances in tree hydraulics highlight the ecological significance of the hydraulic safety margin. *New Phytol.* **203**, 355–358 (2014).
79. Anderegg, W. R. L. et al. Tree mortality predicted from drought-induced vascular damage. *Nat. Geosci.* **8**, 367–371 (2015).
80. McDowell, N. G. et al. Multi-scale predictions of massive conifer mortality due to chronic temperature rise. *Nat. Clim. Change* **6**, 295–300 (2016).
81. Brodribb, T. J., Bowman, D. J. M. S., Nichols, S., Delzon, S. & Burlett, R. Xylem function and growth rate interact to determine recovery rates after exposure to extreme water deficit. *New Phytol.* **188**, 533–542 (2010).
82. Meinzer, F. C., Johnson, D. M., Lachenbruch, B., McCulloh, K. A. & Woodruff, D. R. Xylem hydraulic safety margins in woody plants: coordination of stomatal control of xylem tension with hydraulic capacitance. *Funct. Ecol.* **23**, 922–930 (2009).
83. McDowell, N. G. et al. Evaluating theories of drought-induced vegetation mortality using a multimodel-experiment framework. *New Phytol.* **200**, 304–321 (2013).
84. Gustafson, E. J. & Sturtevant, B. R. Modeling forest mortality caused by drought stress: implications for climate change. *Ecosystems* **16**, 60–74 (2013).
85. Mitchell, P. J. et al. An ecoclimatic framework for evaluating the resilience of vegetation to water deficit. *Glob. Chang. Biol.* **22**, 1677–1689 (2016).
86. O'Sullivan, O. S. et al. Thermal limits of leaf metabolism across biomes. *Glob. Chang. Biol.* **23**, 209–223 (2017).
87. Adams, H. D. et al. Empirical and process-based approaches to climate-induced forest mortality models. *Front. Plant Sci.* **4**, 438 (2013).
88. Christoffersen, B. O. et al. Linking hydraulic traits to tropical forest function in a size-structured and trait-driven model (TFS v.1-Hydro). *Geosci. Model Dev.* **9**, 4227–4255 (2016).
89. Xu, X., Medvigy, D., Powers, J. S., Becknell, J. M. & Guan, K. Diversity in plant hydraulic traits explains seasonal and inter-annual variations of vegetation dynamics in seasonally dry tropical forests. *New Phytol.* **212**, 80–95 (2016).
90. Davi, H. & Cailleret, M. Assessing drought-driven mortality trees with physiological process-based models. *Agric. For. Meteorol.* **232**, 279–290 (2017).
91. Bartlett, M. K., Scoffoni, C. & Sack, L. The determinants of leaf turgor loss point and prediction of drought tolerance of species and biomes: a global meta-analysis. *Ecol. Lett.* **15**, 393–405 (2012).
Data synthesis that links the point at which leaf turgor is lost to drought tolerance in plants.

92. Limousin, J.-M., Longepierre, D., Huc, R. & Rambal, S. Change in hydraulic traits of Mediterranean *Quercus ilex* subjected to long-term throughfall exclusion. *Tree Physiol.* **30**, 1026–1036 (2010).
93. Vilagrosa, A., Bellot, J., Vallejo, V. R. & Gil-Pelegrín, E. Cavitation, stomatal conductance, and leaf dieback in seedlings of two co-occurring Mediterranean shrubs during an intense drought. *J. Exp. Bot.* **54**, 2015–2024 (2003).
94. Dahlin, K. M., Ponte, D. D., Setlock, E. & Nagelkirk, R. Global patterns of drought deciduous phenology in semi-arid and savanna-type ecosystems. *Ecography* **40**, 314–323 (2016).
95. De Kauwe, M. G. et al. Do land surface models need to include differential plant species responses to drought? Examining model predictions across a mesic–xeric gradient in Europe. *Biogeosciences* **12**, 7503–7518 (2015).
96. Aguadé, D., Poyatos, R., Rosas, T. & Martínez-Vilalta, J. Comparative drought responses of *Quercus ilex* L. and *Pinus sylvestris* L. in a montane forest undergoing a vegetation shift. *Forests* **6**, 2505 (2015).
97. Donovan, L., Linton, M. & Richards, J. Predawn plant water potential does not necessarily equilibrate with soil water potential under well-watered conditions. *Oecologia* **129**, 328–335 (2001).
98. Nobel, P. S. & Cui, M. Hydraulic conductances of the soil, the root–soil air gap, and the root: changes for desert succulents in drying soil. *J. Exp. Bot.* **43**, 319–326 (1992).
99. Eller, C. B., Lima, A. L. & Oliveira, R. S. Cloud forest trees with higher foliar water uptake capacity and anisohydric behavior are more vulnerable to drought and climate change. *New Phytol.* **211**, 489–501 (2016).
100. Sinclair, T. R. Model analysis of plant traits leading to prolonged crop survival during severe drought. *Field Crops Res.* **68**, 211–217 (2000).
101. Manzoni, S., Katul, G. & Porporato, A. A dynamical system perspective on plant hydraulic failure. *Wat. Resour. Res.* **50**, 5170–5183 (2014).
102. Gentine, P., Guérin, M., Uriarte, M., McDowell, N. G. & Pockman, W. T. An allometry-based model of the survival strategies of hydraulic failure and carbon starvation. *Ecohydrology* **9**, 529–546 (2016).
103. Waring, R. H. Characteristics of trees predisposed to die. *Bioscience* **37**, 569–574 (1987).
104. Bréda, N., Huc, R., Granier, A. & Dreyer, E. Temperate forest trees and stands under severe drought: a review of ecophysiological responses, adaptation processes and long-term consequences. *Ann. For. Sci.* **63**, 625–644 (2006).
105. De Micco, V., Balzano, A., Wheeler, E. A. & Baas, P. Tyloses and gums: a review of structure, function and occurrence of vessel occlusions. *IAWA J.* **37**, 186–205 (2016).
106. Zeppel, M. J. B. et al. Drought and resprouting plants. *New Phytol.* **206**, 583–589 (2015).
107. Bond, W. J. & Midgley, J. J. Ecology of sprouting in woody plants: the persistence niche. *Trends Ecol. Evol.* **16**, 45–51 (2001).
108. Brodersen, C. R. & McElrone, A. J. Maintenance of xylem network transport capacity: a review of embolism repair in vascular plants. *Front. Plant Sci.* **4**, 108 (2013).
109. Zwieniecki, M. A. & Holbrook, N. M. Confronting Maxwell's demon: biophysics of xylem embolism repair. *Trends Plant Sci.* **14**, 530–534 (2009).
110. Nardini, A., Savi, T., Trifilò, P. & Lo Gullo, M. A. in *Progress in Botany Vol. 79* (eds Cánovas, F. et al.) 197–231 (Springer, Cham, 2017).
111. Cobb, A. R., Choat, B. & Holbrook, N. M. Dynamics of freeze–thaw embolism in *Smilax rotundifolia* (Smilacaceae). *Am. J. Bot.* **94**, 640–649 (2007).
112. Cochard, H., Lemoine, D., Améglio, T. & Granier, A. Mechanisms of xylem recovery from winter embolism in *Fagus sylvatica*. *Tree Physiol.* **21**, 27–33 (2001).
113. Sperry, J. S., Holbrook, N. M., Zimmermann, M. H. & Tyree, M. T. Spring filling of xylem vessels in wild grapevine. *Plant Physiol.* **83**, 414–417 (1987).
114. Kaufmann, I. et al. Functional repair of embolized vessels in maize roots after temporal drought stress, as demonstrated by magnetic resonance imaging. *New Phytol.* **184**, 245–256 (2009).
115. McCully, M. E., Huang, C. X. & Ling, L. E. C. Daily embolism and refilling of xylem vessels in the roots of field-grown maize. *New Phytol.* **138**, 327–342 (1998).
116. Taneda, H. & Sperry, J. S. A case-study of water transport in co-occurring ring- versus diffuse-porous trees: contrasts in water-status, conducting capacity, cavitation and vessel refilling. *Tree Physiol.* **28**, 1641–1651 (2008).
117. Zwieniecki, M. A. & Holbrook, N. M. Diurnal variation in xylem hydraulic conductivity in white ash (*Fraxinus americana* L.), red maple (*Acer rubrum* L.) and red spruce (*Picea rubens* Sarg.). *Plant Cell Environ.* **21**, 1173–1180 (1998).
118. Brodersen, C. R., McElrone, A. J., Choat, B., Matthews, M. A. & Shackel, K. A. The dynamics of embolism repair in xylem: in vivo visualizations using high-resolution computed tomography. *Plant Physiol.* **154**, 1088–1095 (2010). **First study to utilize synchrotron-based imaging methods for non-destructive visualization of xylem function**.
119. Charrier, G. et al. Evidence for hydraulic vulnerability segmentation and lack of xylem refilling under tension. *Plant Physiol.* **172**, 1657–1668 (2016).
120. Clearwater, M. J. & Clark, C. J. In vivo magnetic resonance imaging of xylem vessel contents in woody lianas. *Plant Cell Environ.* **26**, 1205–1214 (2003).
121. Knipfer, T., Brodersen, C. R., Zedan, A., Kluepfel, D. A. & McElrone, A. J. Patterns of drought-induced embolism formation and spread in living walnut saplings visualized using X-ray microtomography. *Tree Physiol.* **35**, 744–755 (2015).
122. Skelton, R. P., Brodribb, T. J., McAdam, S. A. M. & Mitchell, P. J. Gas exchange recovery following natural drought is rapid unless limited by loss of leaf hydraulic conductance: evidence from an evergreen woodland. *New Phytol.* **215**, 1399–1412 (2017).
123. Fukuda, H. Xylogenesis: initiation, progression, and cell death. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 299–325 (1996).
124. Davies, W., Metcalfe, J., Lodge, T. & da Costa, A. R. Plant growth substances and the regulation of growth under drought. *Funct. Plant Biol.* **13**, 105–125 (1986).
125. Liang, E., Balducci, L., Ren, P. & Rossi, S. in *Secondary Xylem Biology: Origins, Functions, and Applications* (eds Kim, Y.S. et al.) 45–58 (Academic, Cambridge, 2016).
126. Hartmann, H. Will a 385 million year-struggle for light become a struggle for water and for carbon? – How trees may cope with more frequent climate change-type drought events. *Glob. Chang. Biol.* **17**, 642–655 (2011).
127. Brunner, I., Herzog, C., Dawes, M. A., Arend, M. & Sperisen, C. How tree roots respond to drought. *Front. Plant Sci.* **6**, 547 (2015).
128. Kozlowski, T. & Pallardy, S. Acclimation and adaptive responses of woody plants to environmental stresses. *Bot. Rev.* **68**, 270–334 (2002).
129. Pagay, V. et al. A microtensiometer capable of measuring water potentials below −10 MPa. *Lab Chip* **14**, 2806–2817 (2014).
130. Luo, Z. et al. Responses of plant water use to a severe summer drought for two subtropical tree species in the central southern China. *J. Hydrol.* **8**, 1–9 (2016).
131. Cochard, H. et al. Methods for measuring plant vulnerability to cavitation: a critical review. *J. Exp. Bot.* **64**, 4779–4791 (2013).
132. Poyatos, R. et al. SAPFLUXNET: towards a global database of sap flow measurements. *Tree Physiol.* **36**, 1449–1455 (2016).
133. McDowell, N. G. et al. Global satellite monitoring of climate-induced vegetation disturbances. *Trends Plant Sci.* **20**, 114–123 (2015).

Acknowledgements We thank S. Stuart, H. Cochard and M. Holbrook for insightful comments and discussion during the preparation of the Review. Micro-computed tomography images included in Fig. 2 were collected during beam-time allocations at the Imaging and Medical beam line (Australian Synchrotron) and TOMCAT beam line (Swiss Light Source). B.C., T.J.B. and B.E.M. acknowledge support from the Australian Research Council (FT130101115; LP140100232; DP170100761). R.L. was supported by a Marie Curie Fellowship (FP7PEOPLE-2013-IOF-624473).

Reviewer information *Nature* thanks B. Engelbrecht, N. G. McDowell and M. Mencuccini for their contribution to the peer review of this work.

Author contributions All authors contributed to writing and planning of the manuscript. B.C., T.J.B. and B.E.M. developed the initial outline and synopsis of the Review. B.C. was responsible for the coordination of the writing of the manuscript. B.C. and C.R.B. prepared figures and the table.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0240-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to B.C.

ARTICLE

https://doi.org/10.1038/s41586-018-0217-9

Targeted therapy in patients with PIK3CA–related overgrowth syndrome

Quitterie Venot[1], Thomas Blanc[1,2,3,21], Smail Hadj Rabia[2,4,5,21], Laureline Berteloot[5,6], Sophia Ladraa[1], Jean-Paul Duong[2,7], Estelle Blanc[8], Simon C. Johnson[9], Clément Hoguin[1], Olivia Boccara[4], Sabine Sarnacki[2,3], Nathalie Boddaert[2,5,6], Stephanie Pannier[2,10], Frank Martinez[11], Sato Magassa[1], Junna Yamaguchi[1], Bertrand Knebelmann[1,2,11], Pierre Merville[12,13], Nicolas Grenier[14], Dominique Joly[1,2,11], Valérie Cormier-Daire[2,5,15], Caroline Michot[2,5,15], Christine Bole-Feysot[5], Arnaud Picard[2,16], Véronique Soupre[16], Stanislas Lyonnet[2,5,15], Jeremy Sadoine[17], Lotfi Slimani[17], Catherine Chaussain[2,17], Cécile Laroche-Raynaud[18], Laurent Guibaud[19], Christine Broissand[20], Jeanne Amiel[2,5,15], Christophe Legendre[1,2,11], Fabiola Terzi[1,2] & Guillaume Canaud[1,2,11]*

CLOVES syndrome (congenital lipomatous overgrowth, vascular malformations, epidermal naevi, scoliosis/skeletal and spinal syndrome) is a genetic disorder that results from somatic, mosaic gain-of-function mutations of the *PIK3CA* gene, and belongs to the spectrum of *PIK3CA*-related overgrowth syndromes (PROS). This rare condition has no specific treatment and a poor survival rate. Here, we describe a postnatal mouse model of PROS/CLOVES that partially recapitulates the human disease, and demonstrate the efficacy of BYL719, an inhibitor of PIK3CA, in preventing and improving organ dysfunction. On the basis of these results, we used BYL719 to treat nineteen patients with PROS. The drug improved the disease symptoms in all patients. Previously intractable vascular tumours became smaller, congestive heart failure was improved, hemihypertrophy was reduced, and scoliosis was attenuated. The treatment was not associated with any substantial side effects. In conclusion, this study provides the first direct evidence supporting PIK3CA inhibition as a promising therapeutic strategy in patients with PROS.

The phosphoinositide-3 kinases (PI(3)Ks) are key lipid kinases that control signalling pathways involved in cell proliferation, motility, survival and metabolism[1]. Class I PI(3)K contains four catalytic isoforms (p110α, p110β, p110δ and p110γ), which carry out non-redundant signalling functions[1]. *PIK3CA* encodes the 110-kD catalytic α-subunit of PI(3)K (p110α), which converts phosphatidylinositol (3,4)-bisphosphate (PtdIns(3,4)P_2) to phosphatidylinositol (3,4,5)-triphosphate (PtdIns(3,4,5)P_3). This leads to the activation of PDK1, which phosphorylates AKT on Thr308[2,3]. AKT, through the phosphorylation and inhibition of tuberous sclerosis complex 2 (TSC2), activates the kinase mechanistic target of rapamycin complex 1 (mTORC1)[4]. PIK3CA is a master regulator of cell growth and other pathways, including the Rho/Rac1 signalling cascade. PROS are caused by post-zygotic mosaic gain-of-function mutations in the *PIK3CA* gene[5], resulting in mosaic PI3KCA activation, which drives various patient phenotypes, ranging from isolated macrodactyly to CLOVES syndrome[6]. Notably, an increasing number of phenotypes has been included in PROS as a result of the identification of *PIK3CA* mutations in previously uncharacterized overgrowth syndromes. Thus, the prevalence of PROS is difficult to estimate owing to the variability of correct clinical diagnoses and the broad phenotypic spectrum of the disorders[6].

There are no animal models of PROS that can be used to understand its physiopathology, and no specific treatments for patients[6,7]. Patients with PROS mainly receive supportive care based on debulking and mutilating surgery, sclerotherapy, and psychological and nutritional support. Inhibition of mTORC seems to improve lymphatic malformations in some patients but is associated with many side effects[5,8].

Several PIKC3A inhibitors are under development as treatments for oncological conditions, where gain-of-function mutations in *PIK3CA* are observed[9]. Among them, BYL719 shows dose- and time-dependent inhibition of the PI3K/AKT pathway in *PIK3CA*-dependent xenograft tumour models[10,11]. As BYL719 is currently under investigation in clinical trials in patients with *PIK3CA*-dependent tumours and has good tolerability[12], we decided to explore its therapeutic potential in PROS.

A mouse model of PROS/CLOVES

We started by developing a mouse model of PROS/CLOVES. To this end, we took advantage of the transgenic mouse strain, *R26StopFLP110**[13,14], which, after breeding with Cre recombinase mice, express a dominant active *PI3KCA* transgene (Extended Data Fig. 1a, b). *R26StopFLP110** mice were crossed with *CAGG-CreER* mice[15] to generate *PIK3CA*$^{CAGG\text{-}CreER}$ animals that ubiquitously express *PIK3CA* upon tamoxifen administration, and EGFP to monitor *Cre* recombination. Three-week-old mice were treated with a single dose of 40 mg kg^{-1} tamoxifen to induce *Cre* recombination[15]. EGFP staining revealed that *Cre* recombination occurred in 25.5 ± 8.3% of splenocytes (Extended Data Fig. 1c). After tamoxifen administration, we observed that the *PIK3CA*$^{CAGG\text{-}CreER}$ mice started to rapidly die with 50% of mortality at

[1]INSERM U1151, Institut Necker Enfants Malades, Paris, France. [2]Université Paris Descartes, Sorbonne Paris Cité, Paris, France. [3]Service de Chirurgie Viscérale Pédiatrique, Hôpital Necker-Enfants Malades, AP-HP, Paris, France. [4]Service de Dermatologie Pédiatrique, Hôpital Necker-Enfants Malades, AP-HP, Paris, France. [5]UMR-1163 Institut Imagine, Hôpital Necker-Enfants Malades, AP-HP, Paris, France. [6]Département de Radiologie Pédiatrique, Hôpital Necker-Enfants Malades, AP-HP, Paris, France. [7]Département d'Anatomopathologie, Hôpital Necker-Enfants Malades, AP-HP, Paris, France. [8]Département de Médecine Nucléaire, Hôpital Marie Lannelongue, Le Plessis Robinsson, France. [9]Department of Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA, USA. [10]Service d'Orthopédie Pédiatrique, Hôpital Necker-Enfants Malades, AP-HP, Paris, France. [11]Service de Néphrologie Transplantation Adultes, Hôpital Necker-Enfants Malades, AP-HP, Paris, France. [12]Service de Néphrologie, Transplantation, Dialyse, Aphérèses, Centre Hospitalier Universitaire Pellegrin, Bordeaux, France. [13]UMR CNRS 5164, Immuno ConcEpT, CNRS, Bordeaux, France. [14]Service d'Imagerie Diagnostique et Interventionnelle de l'Adulte, Centre Hospitalier Universitaire Pellegrin, Bordeaux, France. [15]Service de Génétique Médicale, Hôpital Necker-Enfants Malades, AP-HP, Paris, France. [16]Service de Chirurgie Maxillo-faciale et Chirurgie Plastique, Hôpital Necker-Enfants Malades, AP-HP, Paris, France. [17]Laboratory EA 2496 Orofacial Pathologies, Imaging and Biotherapies, Montrouge, France. [18]Service de Neuropédiatrie, Hôpital de la Mère et de l'Enfant, Limoges, France. [19]Service d'Imagerie Pédiatrique, Hôpital Femme-Mère-Enfant, Bron, France. [20]Pharmacie, Hôpital Necker-Enfants Malades, AP-HP, Paris, France. [21]These authors contributed equally: Thomas Blanc, Smail Hadj Rabia. *e-mail: guillaume.canaud@inserm.fr

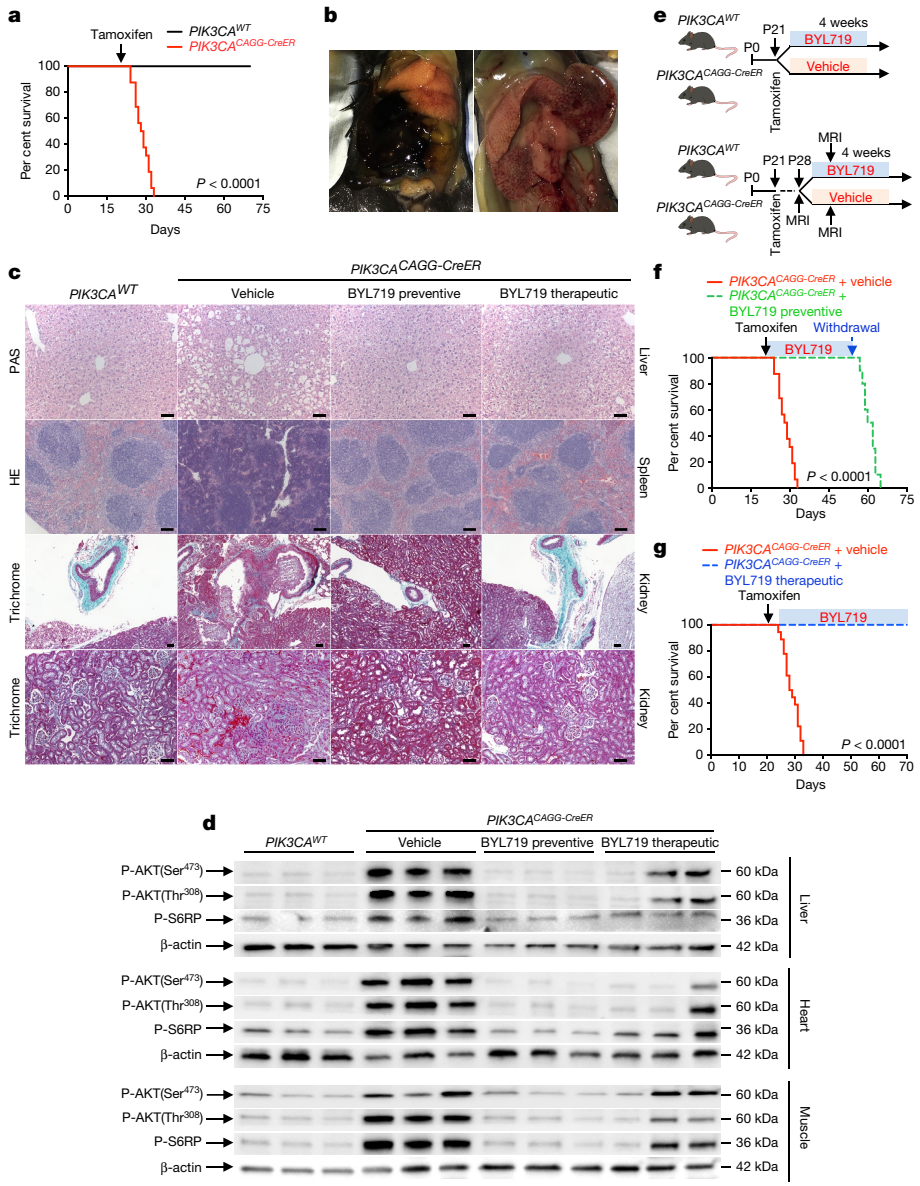


Fig. 1 | Characterization of the PROS mouse model and efficacy of BYL719. a, Kaplan–Meier survival curves of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice ($n = 16$ mice per group) after tamoxifen administration (log-rank test, $P < 0.0001$). **b**, Representative necropsy examination pictures of *PIK3CA*$^{CAGG\text{-}CreER}$ mice. The mice displayed sudden intra-abdominal and spontaneous hepatic haemorrhage ($n = 16$ mice). **c**, Morphology of livers (top), spleens (middle) and kidneys (bottom two rows) from *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice that were treated with or without BYL719 directly after *Cre* induction (preventive) or seven days later (therapeutic) ($n = 8$ mice per group). **d**, Western blot of P-AKT (Ser^{473}), P-AKT (Thr^{308}) and P-S6RP in liver, heart and muscles, respectively, from *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with or without BYL719 directly after *Cre* induction (preventive) or seven days later (therapeutic) ($n = 8$ mice per group). **e**, Experimental design. **f**, Kaplan–Meier survival curves of *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with or without BYL719 after tamoxifen administration ($n = 16$ mice per group). After 40 days, BYL719 treatment was withdrawn (log-rank test, $P < 0.0001$). **g**, Kaplan–Meier survival curves of *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with or without BYL719 ($n = 12$ mice per group) seven days after tamoxifen administration (log-rank test, $P < 0.0001$). Scale bars, 20 μm.

day 9 (Fig. 1a). Death occurred suddenly in most cases, with necropsy revealing intra-abdominal and hepatic haemorrhages (Fig. 1b). Whole body magnetic resonance imaging (MRI) showed scoliosis, vessel abnormalities, kidney cysts, and muscle hypertrophy (Extended Data Fig. 1e, left and middle columns). Histological examination revealed multiple organ abnormalities, including severe liver steatosis with vessel disorganization, loss of spleen microarchitecture integrity, spontaneous haemorrhages and fibrosis of the kidney with aberrant vessels (Fig. 1c, Extended Data Fig. 2a, b). We observed a high number of proliferating cells in all the affected organs of the *PIK3CA*$^{CAGG\text{-}CreER}$ mice, as assed by Ki67 immunostaining (Extended Data Fig. 3a). By contrast, we failed to detect any change in apoptosis (Extended Data Fig. 3b) or senescence (Extended Data Fig. 4a, b). To further characterize the previously reported vessel malformations, we performed CD34 and CD31 immunostaining, which confirmed the presence of severe vessel dilation (Extended Data Fig. 2c). Similarly, podoplanin and LYVE-1 confirmed expansion of lymphatic vessels (Extended Data Fig. 2d, e). We confirmed that mutant p110α (p110*) was expressed in all affected organs (Extended Data Fig. 5a). Notably, we did not detect expression in the brain or lungs (Extended Data Fig. 5b), consistent with the normal histology of these organs (data not shown). Western blot and immunofluorescence studies showed AKT/mTORC pathway activation in all the examined organs (Fig. 1d and Extended Data Figs. 5c–e, 6). In transfected HeLa cells, the *p110** mutant had a stronger effect on AKT/mTORC pathway activation than did the *c.3140A>G* (H1047R) and *c.1633G>A* (E545K) mutations from patients used in previously reported models[16,17] (Extended Data Fig. 7), which may explain the increased phenotypical severity observed in our model.

BYL719 prevents the PROS/CLOVES phenotype

We next studied the effect of BYL719 on *PIK3CA*$^{CAGG\text{-}CreER}$ mice. Initially, BYL719 was administered orally each day starting on the day of *Cre* induction (Fig. 1e). Whereas all placebo-treated *PIK3CA*$^{CAGG\text{-}CreER}$ mice died within 15 days, all BYL719-treated *PIK3CA*$^{CAGG\text{-}CreER}$ mice were alive after 40 days and had an overtly normal appearance (Fig. 1f). Histological examination at 40 days showed that the BYL719 treated mice had preserved tissues (Fig. 1c and Extended Data Fig. 2a) and normal vessels (Extended Data Fig. 2c). Mechanistically, we observed a strong reduction of proliferation in *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with BYL719 (Extended Data Fig. 3a), whereas senescence and apoptosis did not change (Extended Data Figs. 3b, 4a, b). Western blot and immunofluorescence analyses confirmed that PI3KCA pathway activation was effectively inhibited (Fig. 1d and Extended Data Fig. 5c–e, 6). However, interruption of treatment 40 days after *Cre* recombination led to the rapid death of all *PIK3CA*$^{CAGG\text{-}CreER}$ mice (median survival 9.8 days after withdrawal of the drug; Fig. 1f).

Next, we administered either placebo or BYL719 to *PIK3CA*$^{CAGG\text{-}CreER}$ mice seven days after *Cre* induction (Fig. 1e), when tissue abnormalities were already detected by MRI (Extended Data Fig. 1e). BYL719 treatment improved the survival of the *PIK3CA*$^{CAGG\text{-}CreER}$ mice (Fig. 1g). MRI performed 12 days after the start of treatment (19 days after *Cre* induction) showed improvements in scoliosis, muscle hypertrophy, and vessel malformations (Extended Data Fig. 1e). Histological analysis of BYL719-treated mice revealed only minor tissue changes compared to wild-type mice (Fig. 1c, Extended Data Fig. 2a). BYL719 administration strongly reduced cell proliferation in all affected organs (Extended Data Fig. 3a), and western blot and immunofluorescence analyses confirmed PIK3CA pathway inhibition (Fig. 1d, Extended Data Figs. 5c–e, 6).

To try to reproduce more faithfully the lower mosaicism observed in patients[18], we used a single dose of 4 mg kg^{-1} tamoxifen to induce *Cre* recombination. We confirmed that a lower dose of tamoxifen induced a lower rate of mosaicism by fluorescence-activated cell sorting (FACS; 4.1 ± 1.3%, Extended Data Fig. 1d). The mice survived for two months and then died with multiple phenotypic abnormalities. In addition to organomegaly[19], these mice progressively developed asymmetrical overgrowth of extremities, disseminated voluminous tumours and visible subcutaneous vascular abnormalities (Fig. 2a). Histological examination of the tumours revealed the same lesions observed in patients with PROS (lipomatous tumours (Fig. 2b) and severe vascular lesions mixing venous and arterial vessels (Fig. 2c–e). More precisely, we observed ecstatic venous channels with a thin endothelial cell lining, surrounded by sparse, erratically distributed vascular smooth muscle cells and a disorganized extracellular matrix, as is seen in PROS/CLOVES (Fig. 2e). Immunofluorescence and immunochemistry studies confirmed the increase in cell proliferation and activation of the AKT/mTORC pathway (Fig. 2f, Extended Data Fig. 7b).

Once the lesions were clinically visible, we treated the low-dose *PIK3CA*$^{CAGG\text{-}CreER}$ mice with BYL719. Remarkably, the treatment led to the reduction and disappearance of all visible tumours within two weeks, with body weight loss (Fig. 2g, Extended Data Fig. 8a). Computerized tomography (CT) scanning showed a marked and rapid reduction in tumour volume (−83.96 ± 2.81%), with no change in fat content (Extended Data Fig. 8b).

Notably, withdrawal of BYL719 led to the recurrence of tumours and the development of asymmetric extremity hypertrophy within four weeks (Fig. 2h, Extended Data Fig. 7c). These data suggest that continuous administration of BYL719 can relieve PROS symptoms at different stages by inhibiting hyperactive PI3KCA signalling.

BYL719 is more effective than rapamycin

As the mTOR inhibitor rapamycin has shown some efficacy in treating vascular malformation[20], we tested its efficacy in our mouse models. In the first *PIK3CA*$^{CAGG\text{-}CreER}$ mouse model, which received 40 mg kg^{-1} tamoxifen, rapamycin improved survival (Extended Data Fig. 9a). However, it did not improve organ abnormalities such as liver and spleen disorganization with aberrant vascularization (Extended Data Fig. 9b, c). Western blot analysis revealed that rapamycin blocked mTORC1 activity but, as previously reported[21], did not inhibit mTORC2 as assessed by the persistent phosphorylation of AKT on Ser473 in all the examined organs (Extended Data Fig. 9d). Similarly, in the *PIK3CA*$^{CAGG\text{-}CreER}$ mouse model that received 4 mg kg^{-1} tamoxifen, rapamycin treatment did not reduce tumour growth significantly (Extended Data Fig. 9e). In parallel to these studies, we exposed primary mouse fibroblasts from PROS mice to either rapamycin or BYL719. Mutant p110* overexpression led to the phosphorylation of AKT on both Ser473 and Thr308 as well as the phosphorylation of S6RP (Extended Data Fig. 10a, b). By contrast, we could not detect any activation of the p38 and p44/42 MAP kinase pathway. Notably, unlike rapamycin, BYL719 blocked AKT phosphorylation on both Thr308 and Ser473, suggesting that the beneficial effect of BYL719 could be driven by more complete blockage of AKT (Extended Data Fig. 10a, b). These results confirmed that BYL719 could be a good therapeutic option even for patients for whom rapamycin is not an effective option.

BYL719 in two patients with CLOVES syndrome

We administered BYL719 under compassionate care protocols to two patients, one adult and one child, who were suffering from extremely severe clinical manifestations of PROS/CLOVES syndrome with therapeutic failure and life-threatening complications. Patient 1 was a 29-year-old man with *PIK3CA c.3140A>G* (H1047R) mutation (11% mosaicism). He presented left leg hypertrophy, scoliosis, multiple naevi, and extremely severe vascular malformations (Fig. 3a–c). He had previously undergone multiple tumour debulking surgeries and angiographies. He became paraplegic at the age of 20 owing to spinal cord compression and required bladder stenting. He developed progressively worsening, drug-resistant, severe systolic heart failure with a measured cardiac output of 18 l min^{-1} m^{-2} (normal < 3.5 l min^{-1} m^{-2}). Consistently, the brain natriuretic peptide (BNP) level was over 2,500 pg ml^{-1} (normal < 100 pg ml^{-1}) (Fig. 3d). Over five years, the patient received different doses of rapamycin[22] to limit vascular tumour growth progression, but this had no overt effects on his abdominal and dorsal vascular tumours. Finally, the patient developed kidney dysfunction with severe proteinuria. A kidney biopsy revealed glomerular lesions with extensive fibrosis. Such lesions may be the result of heart dysfunction, the administration of rapamycin[5,8], and/or *PIK3CA* mutations in the kidney epithelial cells. CT scan (Fig. 3b) and MRI (Fig. 3e) analyses showed severe vascular malformations, while positron emission tomography (PET) scan analysis was negative. Karnofsky and Eastern cooperative oncology group (ECOG) scores, which indicate general health status[23,24], were 20% and 4, respectively. Owing to the severity of the case and poor prognosis, the physicians, surgeons, and radiologists decided to stop any interventional treatment and to provide the patient with only supportive and palliative care. His estimated life expectancy was less than a few months. One month after rapamycin washout, daily oral BYL719 was started at 250 mg per day, the lowest dose used in clinical trials. We observed a rapid improvement in the general status of the patient, as well as a progressive reduction in tumour size, venous dilations and skin lesions (Fig. 3a). He lost 25 kg in 18 months due to oedema clearance but also to a major reduction in vascular tumour abnormalities (Fig. 3c). His thorax and abdomen circumferences were reduced by 25% and 39%, respectively, over 18 months (Fig. 3c). CT scan and MRI analyses confirmed global vascular tumour shrinkage (72% volume decrease at 18 months), as well as disappearance of subcutaneous infiltration (Fig. 3b, e). The effect of the drug on heart function was remarkable, with complete correction of BNP levels within 4 weeks (Fig. 3d). Cardiac output decreased to 3 l min^{-1} m^{-2}, and his heart size reduced by 25% (Fig. 3b). The left ventricular mass index to body surface area decreased from 250 g m^{-2} to 148 g m^{-2}. Renal function also improved rapidly, with estimated glomerular filtration rate increasing from 33 to 52 ml min^{-1} per 1.73 m^2 (Fig. 3d). The hypertrophy of the left leg was also reduced (Fig. 3a). We

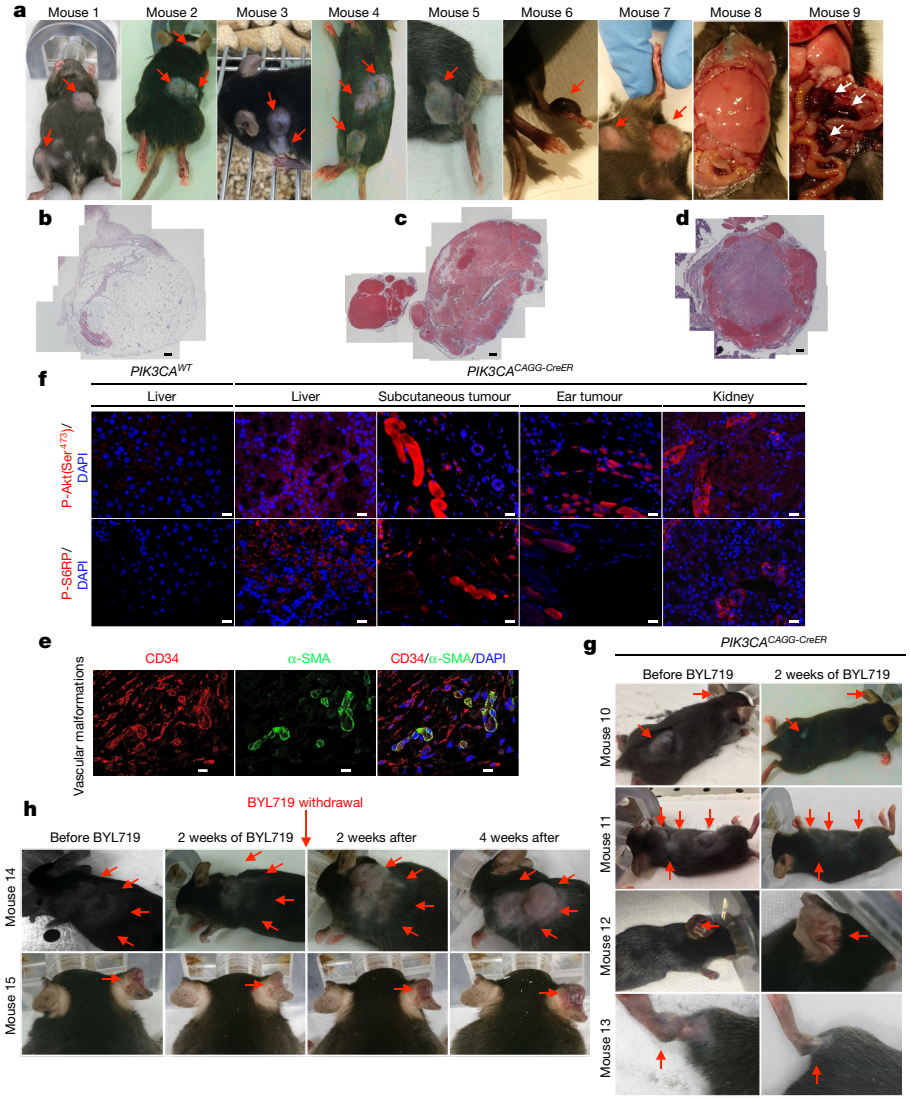


Fig. 2 | Characterization of the second mouse model of PROS and efficacy of BYL719. a, Morphology of *PIK3CA*$^{CAGG\text{-}CreER}$ mice one month after *Cre* induction with a single dose of 4 mg kg^{-1} tamoxifen. The mice displayed multiple visible deforming tumours across the whole body (red arrows). Some of these tumours had huge vascular irregularities. The mice progressively developed hypertrophic extremities, enlarged liver with aberrant vessels, and multiple intraabdominal vascular malformations (pictures representative of nine mice). **b**–**d**, Histopathological examination of the tumours revealed the presence of lipomatous tumours (**b**; $n = 10$ mice examined) and multiple venous malformations composed of ecstatic venous channels with a thin endothelial cell lining, surrounded by sparse, erratically distributed vascular smooth muscle cells and a disorganized extracellular matrix (**c**, **d**; $n = 10$ mice examined). Scale bars, 50 μm. **e**, Coimmunostaining of CD34 and α-smooth muscle cells in venous malformation ($n = 10$ mice examined). **f**, Immunofluorescence staining for P-AKT (Ser^{473}) and P-S6RP in the liver, subcutaneous tumour, ear tumour and kidneys of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice ($n = 10$ mice examined per group). Scale bars, 10 μm. **g**, *PIK3CA*$^{CAGG\text{-}CreER}$ mice were injected with a single dose of 4 mg kg^{-1} tamoxifen and followed for one month after *Cre* induction. Once the tumours reached a certain volume, the mice were treated with BYL719 for two weeks and this led to the disappearance of all tumours ($n = 18$ mice). **h**, *PIK3CA*$^{CAGG\text{-}CreER}$ mice were injected with a single dose of 4 mg kg^{-1} tamoxifen and followed for one month after *Cre* induction. Once the tumours reached a certain volume, the mice were treated with BYL719 for two weeks and this led to the disappearance of all tumours ($n = 7$ mice). Then, BYL719 treatment was withdrawal and we observed the recurrence of tumours and vascular malformations in *PIK3CA*$^{CAGG\text{-}CreER}$ mice within the next four weeks.

observed a marked improvement in skin hypertrophy, with a change in nevus coloration and a reduction in ear size (Fig. 3a). Finally, after 6 months of BYL719 treatment, the patient began to partially gain bladder function, with an improvement in saddle anaesthesia. MRI revealed a 60% reduction in the size of the venous malformation that was compressing his spinal cord. The patient's Karnofsky and ECOG scores increased to 80% and 1, respectively. After eighteen months of treatment, the patient had no side effects except hyperglycaemia which was well controlled with nutrition therapy. BYL719 leads to peripheral insulin resistance[11] and hyperglycaemia was one of the most common drug-related adverse events in the phase I trial[12]. The patient is currently still under BYL719 treatment.

Patient 2 was a nine-year-old girl with a PROS/CLOVES syndrome diagnosed two years before this study and a gain of function *PIK3CA c.3140A>G* (H1047R) mosaicism mutation (12% mosaicism) confirmed by skin biopsy. She had scoliosis, left leg hypertrophy, vascular malformations, and hypertrophic back muscles (Fig. 4a). She also had a voluminous cystic lymphangioma involving the left kidney and the gastrointestinal tract. A PET scan showed metabolic fixation in the thymus, back muscles, and left leg muscles (Fig. 4d). Her Karnofsky and ECOG scores were 50% and 3, respectively. BYL719 was started at the lowest available dosage: 50 mg per day. As in the first patient, we observed a marked clinical improvement rapidly after the onset of treatment (Fig. 4a). The patient reported improved comfort and we observed a

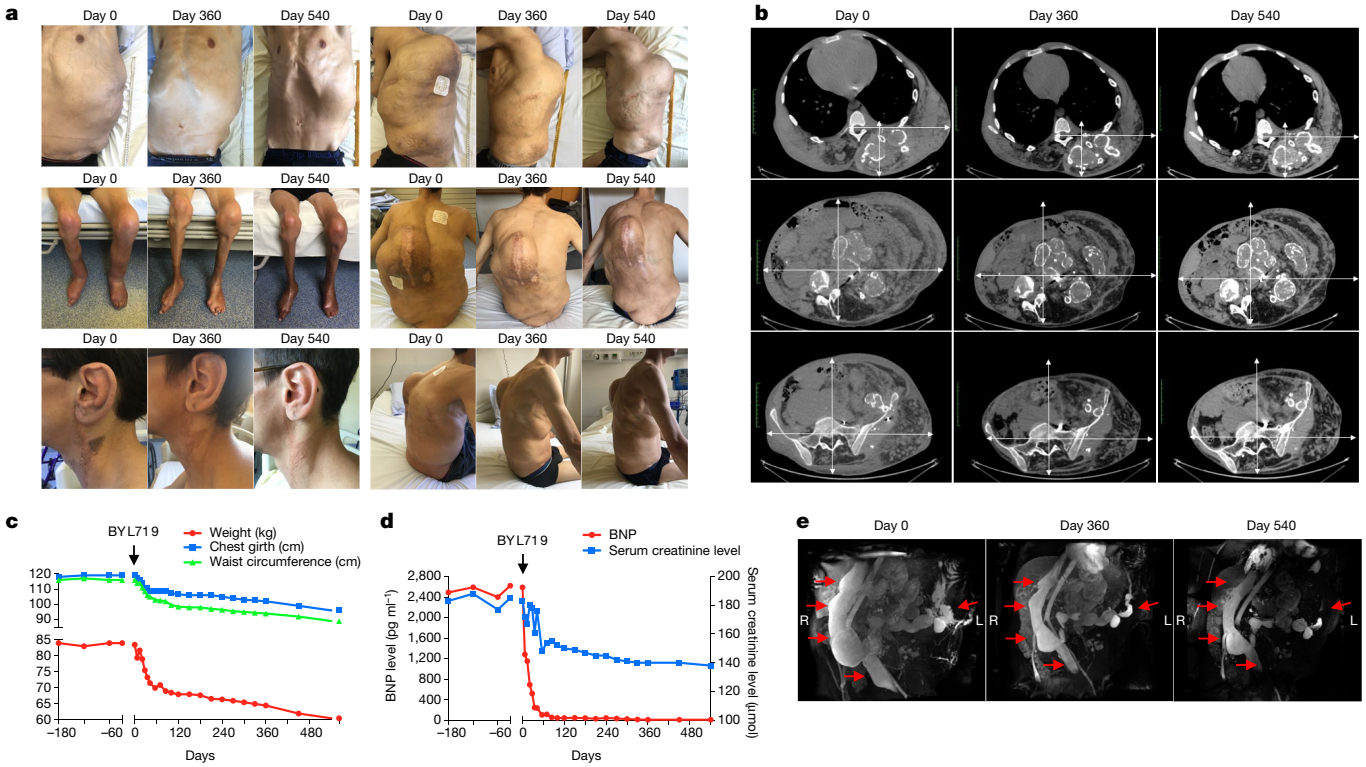


Fig. 3 | Efficacy of BYL719 treatment in an adult patient with severe CLOVES syndrome. a, Representative pictures of patient 1 before and 360 and 540 days after onset of BYL719 treatment. **b**, CT scans of patient 1 before and 360 and 540 days after onset of BYL719 treatment. The magnifications are the same. **c**, Clinical parameters (weight, chest girth and waist circumference) of patient 1 before and after onset of BYL719 treatment. **d**, Serum BNP and creatinine levels before and after onset of BYL719 treatment. **e**, Three-dimensional MRI-based reconstructions of the intra-abdominal vessels before and after onset of BYL719 treatment. The total volume of the venous malformation was reduced by up to 72% after 540 days of treatment.

24% reduction in her left leg volume, reduced left foot volume, and a 14% reduction in her abdominal circumference (Fig. 4b) after 12 months. The hypertrophic muscles in her back also shrank rapidly and, unexpectedly, her scoliosis was reversed after 6 months without any other intervention (Fig. 4c). More importantly, MRI showed a 40%, 54% and 71% reduction in intra-abdominal tumour volume after 4, 6 and 12 months of treatment, respectively (Fig. 4e). A PET scan performed at the same time points confirmed the marked reduction in metabolic activity of the affected tissues (Fig. 4d). The patient's Karnofsky and ECOG scores were greatly improved, to 100% and 0, respectively. Her height had increased

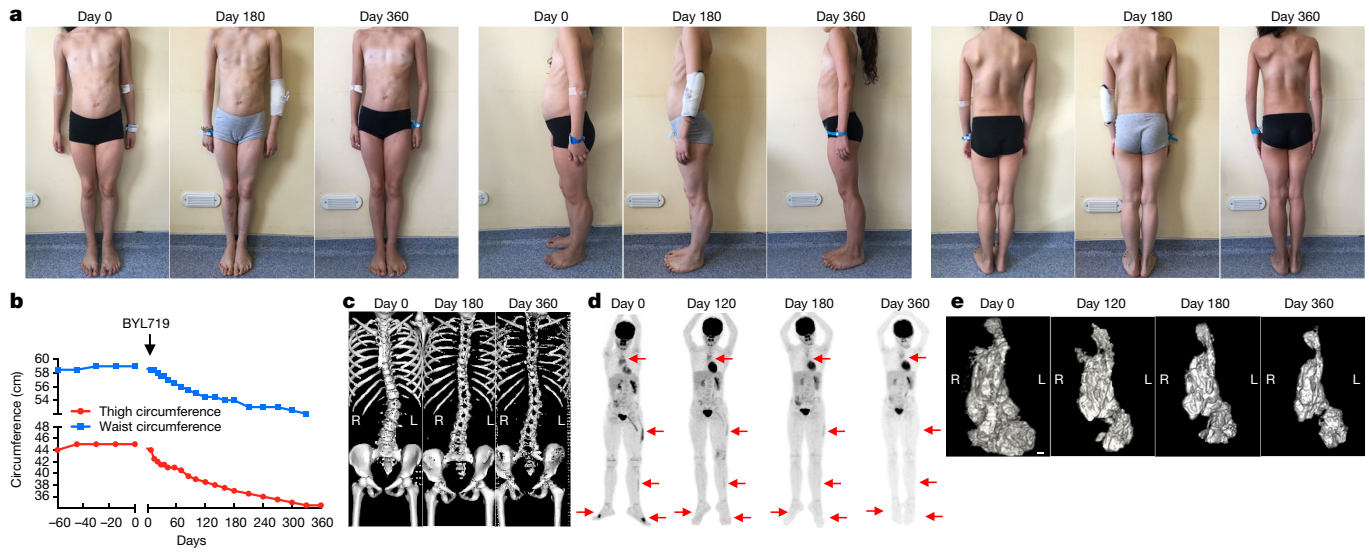


Fig. 4 | Efficacy of BYL719 treatment in a child patient with severe CLOVES syndrome. a, Representative pictures of patient 2 before and 180 and 360 days after onset of BYL719 treatment. **b**, Clinical parameters (thigh and waist circumferences) of patient 2 before and after onset of BYL719 treatment. **c**, Three-dimensional CT scan reconstruction of the spine before and 180 and 360 days after onset of BYL719 treatment. **d**, PET scan images of patient 2 before and 120, 180 and 360 days after onset of BYL719 treatment. Arrows indicate the hypermetabolic tissues before and during treatment. **e**, Three-dimensional MRI-based reconstruction of the intra-abdominal tumour before and 120, 180 and 360 days after onset of BYL719 treatment. The tumour volume was reduced by up to 71% after 360 days of treatment. Scale bar, 1 cm.

Fig. 5 | Efficacy of BYL719 treatment in patients with PROS. Images of patients 1–9 before and after six months of BYL719 treatment. Patient 1 is a four-year-old girl suffering from severe vascular malformations involving the right arm and chest with permanent pain. After six months of treatment we saw a marked improvement in all vascular malformations as well as the scoliosis. Patient 2 is a 4-year-old girl with scoliosis and hypertrophic left buttock who had already undergone left foot partial amputation. After six months of treatment we saw an improvement in the scoliosis and reduction of the hypertrophic lesion. Patient 3 is a 5-year-old girl with CLOVES syndrome and chronic gastrointestinal bleeding. After six months of treatment, chronic bleeding stopped. Patient 4 is a 5-year-old girl with left hemifacial hyperplasia that was progressing despite multiple surgeries. After six months of treatment we saw for the first time an improvement in the hypertrophy (not shown for reasons of confidentiality). Patient 5 is a 6-year-old boy with CLOVES syndrome. The treatment led to a reduction in lipomatous tumours and scoliosis. Patient 6 is a 10-year-old girl with CLOVES syndrome and a severe lipomatous tumour on her back. BYL719 led to a marked improvement in the scoliosis and shrinkage of the tumour. Patient 7 is an 11-year-old boy with CLOVES syndrome, chronic gastrointestinal bleeding and severe dyspnea. Treatment improved all symptoms and the bleeding stopped. Patient 8 is an 11-year-old girl with CLOVES syndrome and severe dyspnea. Treatment improved the subcutaneous lipoma as well as dyspnea. Patient 9 is a 13-year-old girl with CLOVES syndrome, splenomegaly and severe vascular malformations involving the left kidney and limbs. Vascular malformations were markedly improved and the size of the spleen reduced after treatment.

by 0.6 cm per month during the 6 months before BYL719 treatment began, and growth remained stable at 0.6 cm per month during the 12-month treatment period. Her glycaemia remained normal and her glycated haemoglobin (HbA1c) level, which was 5.5% before BYL719 treatment, was 5.1, 4.9 and 5.0% after 3, 6 and 12 months of treatment, respectively. The patient is still receiving BYL719 treatment.

BYL719 in a cohort of PROS patients

On the basis of these results, we were authorized to administer BYL719 to 17 additional patients with PROS who had life-threatening complications and/or were scheduled for debulking surgery (Fig. 5 and Extended Data Fig. 11a). The demographical characteristics of the patients are listed in Supplementary Table 1. All patients had documented gain-of-function mutations in the *PIK3CA* gene, with ten different mutations (Extended Data Fig. 11b and Supplementary Table 1). All patients underwent clinical examination, laboratory evaluation, PET and MRI scans before BYL719 administration and 3 and 6 months after. Fourteen children were started at the lowest available dosage, 50 mg per day and three adults at 250 mg per day. Among the 17 patients, six had been diagnosed with CLOVES syndrome, two with megalencephaly–capillary malformation (MCAP), and nine with localized overgrowth syndrome (back, limbs, face, arms) including three patients with abdominal or chest vascular tumours. In addition, each patient had peculiar clinical features, and twelve patients had undergone previous debulking surgery (Fig. 5 and Extended Data Fig. 11a). Eight patients had received previous rapamycin treatment for18 months without clinical or radiological improvement.

After BYL719, all patients showed substantial clinical improvement (Supplementary Table 1). All patients described reduced tiredness. The mean circumference of the clinical target lesions decreased by 12.6 ± 3.8 and $16.3 \pm 3.9\%$ after 3 and 6 months of BYL719 treatment, respectively (Supplementary Table 2 and Extended Data Fig. 11c). All patients demonstrated an improvement in skin capillary abnormalities and naevi became thinner. The two opioid-dependent patients who were confined to bed could stop morphine, and after 2 months of treatment were able to walk without help; tiredness improved, and haematuria disappeared in patient 11. Chronic gastrointestinal bleeding stopped in three patients, associated with the correction of disseminated intravascular coagulation in patients 7 and 9, which led to the cessation of heparin treatment. We also observed an improvement in chronic palpebral cellulitis in patient 14 and we were able to progressively taper down this patient's steroid treatment. All patients demonstrated a clinical improvement in scoliosis and the surgical corset was removed from patient 6. In the two patients with MCAP (patients 13 and 14), we observed an improvement in cognitive function and behaviour. After 6 months of treatment, all patients were still alive and no surgery had been performed.

Introduction of BYL719 initially led to weight loss with a maximum after two months of treatment ($-1.7 \pm 3.1\%$). However, after five months of treatment, body weight came back to baseline and patients started to gain weight (Extended Data Fig. 11d). Notably, four obese patients showed a marked reduction in body weight ($-9.7 \pm 8.2\%$; Extended Data Fig. 11e). This reduction was not associated with a reduction in food intake. MRI demonstrated an notable reduction in subcutaneous fat in these patients. The growth of the children[4] was not affected by the drug during the six months of treatment (Extended Data Fig. 12a).

In addition to the clinical improvement, we observed a radiological response in all patients. After 90 and 180 days of therapy, the mean volume of the target lesions had decreased by 27.2 ± 14.6 and 37.8 ± 16.3%, respectively (Extended Data Fig. 12b–e). All target lesions responded to treatment. Notably, in the two patients with MCAP, MRI revealed an improvement in cerebral perfusion (Extended Data Fig. 12f). Finally, 5 of the 17 patients had hypermetabolic activity as assessed by PET scan. Remarkably, after 90 days of treatment we observed a drastic reduction in the metabolic activity of affected tissues (Extended Data Fig. 12g).

Although our follow-up period was short, we observed a low rate of side effects using this drug. Indeed, we noticed only discrete mouth ulcerations in three patients during one week that disappeared spontaneously (grade I). BYL719 was not associated with organ toxicity during this period (as assessed by heart, liver, kidney function and blood testing). Transient hyperglycaemia occurred in one obese patient (patient 15). Hyperglycaemia was well controlled with diet restriction only. Patient 17 was diabetic before BYL719 introduction and needed to increase the posology of oral antidiabetic drugs. Glycated haemoglobin remained normal during treatment in all patients (data not shown). All of the patients are still receiving BYL719 treatment.

Discussion

We have described a mouse model of PROS/CLOVES that mimics elements of the human phenotype by allowing titration of mosaicism; we have shown that BYL719 treatment rescues the PROS phenotype in the mouse model more efficaciously than rapamycin; and we have demonstrated that BYL719 is clinically effective, with a promising safety profile, in patients with PROS regardless of the type of *PIK3CA* mutation. This study provides the first direct evidence that PIK3CA inhibition is a robust and effective therapeutic strategy that can markedly improve the outcome of PROS patients.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0217-9.

Received: 1 March 2017; Accepted: 16 May 2018;
Published online: 13 June 2018

1. Vanhaesebroeck, B., Stephens, L. & Hawkins, P. PI3K signalling: the path to discovery and understanding. *Nat. Rev. Mol. Cell Biol.* **13**, 195–203 (2012).
2. Hiles, I. D. et al. Phosphatidylinositol 3-kinase: structure and expression of the 110 kd catalytic subunit. *Cell* **70**, 419–429 (1992).
3. Stephens, L. et al. Protein kinase B kinases that mediate phosphatidylinositol 3,4,5-trisphosphate-dependent activation of protein kinase B. *Science* **279**, 710–714 (1998).
4. Manning, B. D. & Toker, A. AKT/PKB signaling: navigating the network. *Cell* **169**, 381–405 (2017).
5. Keppler-Noreuil, K. M. et al. PIK3CA-related overgrowth spectrum (PROS): diagnostic and testing eligibility criteria, differential diagnosis, and evaluation. *Am. J. Med. Genet. A.* **167A**, 287–295 (2015).
6. Keppler-Noreuil, K. M. et al. Clinical delineation and natural history of the PIK3CA-related overgrowth spectrum. *Am. J. Med. Genet. A.* **164A**, 1713–1733 (2014).
7. Kurek, K. C. et al. Somatic mosaic activating mutations in PIK3CA cause CLOVES syndrome. *Am. J. Hum. Genet.* **90**, 1108–1115 (2012).
8. Canaud, G. et al. AKT2 is essential to maintain podocyte viability and function during chronic kidney disease. *Nat. Med.* **19**, 1288–1296 (2013).
9. Engelman, J. A. Targeting PI3K signalling in cancer: opportunities, challenges and limitations. *Nat. Rev. Cancer* **9**, 550–562 (2009).
10. Furet, P. et al. Discovery of NVP-BYL719 a potent and selective phosphatidylinositol-3 kinase alpha inhibitor selected for clinical evaluation. *Bioorg. Med. Chem. Lett.* **23**, 3741–3748 (2013).
11. Fritsch, C. et al. Characterization of the novel and specific PI3Kα inhibitor NVP-BYL719 and development of the patient stratification strategy for clinical trials. *Mol. Cancer Ther.* **13**, 1117–1129 (2014).
12. Mayer, I. A. et al. A phase Ib study of alpelisib (BYL719), a PI3Kα-specific inhibitor, with letrozole in ER+/HER2− metastatic breast cancer. *Clin. Cancer Res.* **23**, 26–34 (2017).
13. Srinivasan, L. et al. PI3 kinase signals BCR-dependent mature B cell survival. *Cell* **139**, 573–586 (2009).
14. Klippel, A. et al. Membrane localization of phosphatidylinositol 3-kinase is sufficient to activate multiple signal-transducing kinase pathways. *Mol. Cell. Biol.* **16**, 4117–4127 (1996).
15. Hayashi, S. & McMahon, A. P. Efficient recombination in diverse tissues by a tamoxifen-inducible form of Cre: a tool for temporally regulated gene activation/inactivation in the mouse. *Dev. Biol.* **244**, 305–318 (2002).
16. Castillo, S. D. et al. Somatic activating mutations in Pik3ca cause sporadic venous malformations in mice and humans. *Sci. Transl. Med.* **8**, 332ra43 (2016).
17. Hare, L. M. et al. Heterozygous expression of the oncogenic Pik3ca(H1047R) mutation during murine development results in fatal embryonic and extraembryonic defects. *Dev. Biol.* **404**, 14–26 (2015).
18. Mirzaa, G. et al. *PIK3CA*-associated developmental disorders exhibit distinct classes of mutations with variable expression and tissue distribution. *JCI Insight* **1**, e87623 (2016).
19. Kinross, K. M. et al. Ubiquitous expression of the Pik3caH1047R mutation promotes hypoglycemia, hypoinsulinemia, and organomegaly. *FASEB J.* **29**, 1426–1434 (2015).
20. Hammill, A. M. et al. Sirolimus for the treatment of complicated vascular anomalies in children. *Pediatr. Blood Cancer* **57**, 1018–1024 (2011).
21. O'Reilly, K. E. et al. mTOR inhibition induces upstream receptor tyrosine kinase signaling and activates Akt. *Cancer Res.* **66**, 1500–1508 (2006).
22. Heitman, J., Movva, N. R. & Hall, M. N. Targets for cell cycle arrest by the immunosuppressant rapamycin in yeast. *Science* **253**, 905–909 (1991).
23. Velarde-Jurado, E. & Avila-Figueroa, C. [Evaluation of the quality of life]. *Salud Publica Mex.* **44**, 349–361 (2002).
24. Oken, M. M. et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am. J. Clin. Oncol.* **5**, 649–655 (1982).

Acknowledgements This project received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program grants, (STG-2015) 679254 and (PoC-2016) 737546 (both awarded to G.C.). This work was also supported by the Emmanuel BOUSSARD Foundation, the DAY SOLVAY Foundation, Fondation TOURRE, Fondation Simone et Cino Del Duca, INSERM, Assistance Publique—Hôpitaux de Paris and the University of Paris, Descartes. We would like to thank the two patients and their families. We also thank C. Semaille from the ANSM for help and advice; S. Bisot-Locard (Novartis) for the BYL719; G. Autret for performing the mouse MRIs; S. Berissi and colleagues at the Plateforme d'histologie et morphologie du petit animal, INEM, Paris; S. Principe for administrative management; A. Klippel for advice and for providing plasmids encoding *Myr-p110*-myc* and *Myr-p110*KR-myc* constructs; K. Rajewsky for advice; and the radiological team from the Centre d'Imagerie de Franconville and in particular M. Canaud and A. Scemama for their help.

Reviewer information *Nature* thanks E. Baselga, W. Dobyns, R. Semple and B. Vanhaesebroeck for their contribution to the peer review of this work.

Author contributions Q.V. performed the experiments and analysed the data. T.B., S.H.R., O.B., S.S., S.P., F.M., B.K., P.M., N.G., D.J., V.C.-D., C.M., A.P., S.C.J., V.S., S.Ly., C.L.-R., L.G., C.B., J.A. and C.L. followed the patients and analysed the data. L.B. and N.B. performed the patient CT scans and MRIs and analysed the data. E.B. performed the PET scans and analysed the data. S.La., S.M, J.Y. and S.C.J. performed some in vivo and in vitro experiments. C.H. performed some mouse experiments. J.-P.D. analysed all the histological findings. C.B.-F performed *PIK3CA* genotyping in patients. J.S., L.S. and C.C. performed and analysed the mouse CT scans. F.T. was involved in data analysis and helped to write the paper. G.C. followed the patients, provided the conceptual framework, designed the study, supervised the project, and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0217-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0217-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to G.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mice. For this study, we interbred homozygous *R26Stop^{FL}P110** (Stock# 012343) and heterozygous *CAGGCre-ER^{TM}* (Stock# 004682) mice on the C57BL/6 background obtained from Jackson Laboratories. We obtained *R26Stop^{FL}P110*$^{+/-}$* × *CAGGCre-ER^{TM+}* mice (referred to here as *PIK3CA$^{CAGG\text{-}CreER}$* mice) and *R26Stop^{FL}P110*$^{+/+}$* × *CAGGCre-ER^{TM-}* mice (referred to here as *PIK3CA^{WT}* mice). We used the p110* construct[14] (Extended Data Fig. 9). The p110* protein is a constitutively active chimaera that contains the iSH2 domain of p85 fused to the N terminus of p110 via a flexible glycine-linker. To generate tissue-specific p110*-transgenic mice, a cloned loxP-flanked neoR-stop cassette was inserted into a modified version of pROSA26-1 followed by the cDNA encoding p110* and then a frt-flanked IRES-EGFP cassette and a bovine polyadenylation sequence (*R26StopFLP110**)[13].

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Animals were fed ad libitum and housed at a constant ambient temperature in a 12-h light cycle. Animal procedures were approved by the 'Services Vétérinaires de la Préfecture de Police de Paris' Departmental Director and by the ethical committee of the Paris Descartes University. All appropriate procedures were followed to ensure animal welfare. In the first study, a single dose of tamoxifen (40 mg kg^{-1}) was administered by oral gavage when the mice were 21 days old. For the survival studies, the mice were followed daily after tamoxifen gavage (*PIK3CA^{WT}* $n = 16$ and *PIK3CA$^{CAGG\text{-}CreER}$* $n = 16$). For the preventive studies, the mice were treated with BYL719 (MedChem Express; 50 mg kg^{-1} in 0.5% carboxymethylcellulose (Sigma Aldrich), daily p.o.) ($n = 16$) or vehicle (0.5% carboxymethylcellulose (Sigma Aldrich), daily p.o.) ($n = 16$). For the therapeutic studies, the mice were treated with BYL719 (50 mg kg^{-1} in 0.5% carboxymethylcellulose (Sigma Aldrich), daily p.o.) ($n = 12$) or vehicle (0.5% carboxymethylcellulose (Sigma Aldrich), daily p.o.) ($n = 12$). Treatment was started at the time of tamoxifen gavage for the preventive study, or seven days later for the therapeutic study.

In the second study, a single dose of tamoxifen (4 mg kg^{-1}) was administered by oral gavage when the mice were 21 days old (*PIK3CA$^{CAGG\text{-}CreER}$* $n = 28$). Ten mice were killed at approximately 1 month after tamoxifen gavage when tumours reached a certain volume, for tissue examination. Once gross morphological abnormalities were visible, eighteen mice were treated with BYL719 for 15 days and treatment was stopped once macroscopic lesions disappeared to follow their phenotype.

In the third study, a single dose of tamoxifen (40 mg kg^{-1}) was administered through oral gavage when the mice were 21 days old and the mice were then treated with daily intraperitoneal injection of rapamycin (MedChem Express, ref#HY-10219) (*PIK3CA$^{CAGG\text{-}CreER}$* $n = 7$) or vehicle (*PIK3CA$^{CAGG\text{-}CreER}$* $n = 7$) for 30 days. Rapamycin was dissolved to a final concentration of 0.5 mg ml^{-1} in absolute ethanol dissolved further in 5% polyethylene glycol (PEG-400) and 5% Tween 80 in PBS and used for intraperitoneal injection at 4 mg kg^{-1}. All mice treated with rapamycin or vehicle were then killed for tissue examination.

In the fourth study, a single dose of tamoxifen (4 mg kg^{-1}) was administered by oral gavage when the mice were at 21 days old (*PIK3CA$^{CAGG\text{-}CreER}$* $n = 6$). Once gross morphological abnormalities were visible (approximately 30 days later), all the mice were treated with daily intraperitoneal injection of rapamycin (MedChem Express, ref#HY-10219) for 30 days. Rapamycin was dissolved to a final concentration of 0.5 mg ml^{-1} in absolute ethanol dissolved further in 5% polyethylene glycol (PEG-400) and 5% Tween 80 in PBS and used for intraperitoneal injection at 4 mg kg^{-1}.

Cell cultures, plasmids and transfection. HeLa cells were grown at 37 °C in Dulbecco's modified Eagle's medium (DMEM) containing 10% fetal bovine serum (Sigma). Transient transfections were performed with cDNAs encoding wild-type *PIK3CA* (Addgene Inc., ID 16643), *PIK3CA* (Exon20 H1047R) (Addgene Inc., ID 16639) and *PIK3CA* (Exon9 E545K) (Addgene Inc., ID 16642) constructs and *Myr-p110*-myc* and *Myr-p110*KR-myc* constructs[14] using the Lipofectamine 2000 transfection reagent following the manufacturer's instructions. Cells were lysed in RIPA buffer for protein extraction 48 h after transfection and 12 h after starvation. Each experiment was performed in duplicate and repeated at least four times.

Skin samples were collected from the *PIK3CA^{WT}* and *PIK3CA$^{CAGG\text{-}CreER}$* mice using standard methods[25]. To generate dermal fibroblast cultures, skin samples were minced and incubated at room temperature in 0.05% trypsin-EDTA (ThermoFisher) solution for 30 min with gentle shaking. Cells were collected by centrifuging at 700*g* for 10 min, resuspended in cell culture medium containing 25% FBS, and plated onto 24-well plates to establish lines. Fibroblast cultures were grown and maintained in 1× MEM (Corning) supplemented with 25% FBS and penicillin/streptomycin (Corning) to a final concentration of 100 IU penicillin and 500 μg ml^{-1} streptomycin.

Cells at similar population doubling were plated 1:4 from confluent cultures and allowed to grow until ~80% confluent. Then, medium was replaced with medium containing 4-OH tamoxifen (1 μM) (Sigma) for 48 h to activate the *Cre* recombinase. Cells were then starved for 12 h and exposed for 24 h to either BYL719 (0.1 or 1.0 μmol l^{-1}; MedChem Express) or rapamycin (10 or 100 ng ml^{-1}; Sigma) or an equal volume of DMSO. Cells were rinsed with 1× PBS. Protein lysates were collected by directly adding 1× RIPA buffer containing protease and phosphatase inhibitors. Each experiment was performed in duplicate and repeated at least four times

Morphological analysis. Mouse tissues were fixed in 4% paraformaldehyde and paraffin embedded. 4-μm liver sections were stained with periodic acid-Schiff (PAS), 4-μm liver or spleen sections were stained with haematoxylin and eosin (HE), and 4-μm kidney sections were stained with Masson's trichrome. Four-μm liver frozen sections were stained with Oil red O (Sigma Aldrich, ref# 00625).

Immunohistochemistry and immunofluorescence. Paraffin-embedded kidney sections (4 μm) were incubated with anti-P-AKT (Ser^{473}) antibody (Cell Signaling Technology, ref#4051 dilution 1:100), anti-P-S6RP antibody (Cell Signaling Technology, ref# 5364, dilution 1:100), anti-α-smooth muscle cell antibody (Sigma Aldrich, ref# A5228, dilution 1:100), anti-CD34 antibody (eBioscience, ref# 14-0341, dilution 1:100), anti-CD31 antibody (Dianova, ref# Dia-310, dilution 1:30) or anti-podoplanin antibody (Agilent, ref#M3619, dilution 1:50). Immunofluorescence studies were analysed using a Zeiss LSM 700 confocal microscope.

Proliferation assay. Proliferative cells were detected in tissues using Ki67 immunostaining (anti-Ki67 antibody, Thermo Fisher Scientific, ref# RM-9106-S1, dilution 1:100). 4-μm kidney sections were incubated with anti-Ki67 antibody, followed by a biotinylated mouse antibody (Vector) at 1:400 and a HRP-labelled streptavidin at 1:1,000. Staining was revealed by DAB. The proliferation index (PI) per tissue was calculated as the number of Ki67–positive nuclei for the total number of nuclei in 10 randomly selected fields (magnification ×400).

Apoptosis assay. Apoptosis was detected in 4-μm sections of paraffin-embedded tissues (liver, spleen and heart) by TUNEL assay using the in situ Cell Death Detection kit (Roche) according to the manufacturer's protocol. The number of apoptotic cells was determined as the number of TUNEL-positive nuclei per field in 10 randomly selected fields (magnification ×400).

β-galactosidase. Fresh tissues were briefly rinsed in PBS and then washed three times in PBS containing 5 mM EGTA, 2 mM $MgCl_2$, 0.02% nonidet and 0.01% Na desoxycholate. Samples were rinsed in PBS and then fixed in 20% formaldehyde/2% glutaraldehyde for 45 min. Tissues were then immersed overnight in β-galactosidase staining solution. After rinsing in PBS, livers were fixed in 4% formaldehyde and paraffin embedded. 4-μm tissue sections were counterstained with haematoxylin and eosin. We then counted the number of cells with β-galactosidase staining per field in 10 randomly selected fields (magnification ×400).

mRNA analysis. mRNAs were quantified in mouse tissues by quantitative PCR with reverse transcriptase (RT–PCR) using an ABI PRISM 7700 Sequence Detection system (Applied Biosystems). Primers were as follows: p16 (fwd) 5′-GGCCAATCCCAAGAGCAGAG-3′ and (rev) 5′-GCCACATGCTAGACACGCTA-3′; RPL13 (fwd) 5′-CTCATCCTGTTCCCCAGGAA-3′ and (rev) 5′-GGGTGGCCAGCTTAAGTTCTT-3′. RPL13 was used as the normalization control as previously described[26].

Western blot. Western blots were performed as previously described[27]. In brief, protein extracts from the liver, muscles, heart, kidneys and fibroblasts were resolved by SDS–PAGE before being transferred onto the appropriate membrane and incubated with anti-P-AKT (Ser^{473}) antibody (Cell Signaling Technology, ref# 4060, dilution 1:1,000), anti-P-AKT (Thr^{308}) antibody (Cell Signaling Technology, ref# 13038, dilution 1:1,000), anti-AKT antibody (Cell Signaling Technology, ref# 9272, dilution 1:1,000), anti-P-S6RP antibody (Cell Signaling Technology, ref# 5364, dilution 1:1,000), anti-p110α (Cell Signaling Technology, ref# 4249, dilution 1:1,000), anti-LYVE-1 (R&D Systems, ref# AF2125, dilution 1:1,000), anti-P-p44/42 antibody (Cell Signaling Technology, ref# 4370, dilution 1:1,000), anti-p44/42 (Thr^{202}/Tyr^{204}) antibody (Cell Signaling Technology, ref# 9102, dilution 1:1,000), anti-P-p38 (Thr^{180}/Tyr^{182}) antibody (Cell Signaling Technology, ref# 9216, dilution 1:1,000), anti-p38 antibody (Cell Signaling Technology, ref# 8690, dilution 1:1,000), anti-GFP antibody (Abcam, ref# ab13970, dilution 1:1,000), anti-GAPDH (Merck Millipore, ref#374, dilution 1:1,000) or anti-β-actin antibody (Sigma-Aldrich, ref#A2228, dilution 1:1,000), followed by the appropriate peroxidase-conjugated secondary antibody (dilution 1:10,000). Chemiluminescence was acquired using a Fusion FX7 camera (Vilbert Lourmat) and densitometry was performed using the Bio1D software (Certain Tech).

CT scan and MRI evaluation. High resolution micro-CT scans were performed at the Plateforme PIV, EA2496—Pathologie, Imagerie et Biothérapies orofaciales, Dental school Paris V. In brief, 4-week-old female *PIK3CA^{WT}* and *PIK3CA$^{CAGG\text{-}CreER}$* mice induced with 40 mg kg^{-1} tamoxifen received either vehicle ($n = 3$ *PIK3CA^{WT}* mice) or BYL719 ($n = 3$ *PIK3CA^{WT}* mice, $n = 3$ *PIK3CA$^{CAGG\text{-}CreER}$* mice) for 14 days. Additionally, *PIK3CA$^{CAGG\text{-}CreER}$* mice ($n = 3$ mice) received

a single dose of 4 mg kg^{-1} tamoxifen one month earlier in order to induce the development of tumours and hypertrophic extremities. High-resolution micro-CT scans were performed under general anaesthesia before treatment and 7 and 14 days later.

MRI scans were performed at the Plateforme IRM, INSERM U970, Centre de recherche cardiovasculaire de Paris. In brief, MRI scans were performed under general anaesthesia in 4-week-old female *PIK3CA*^{WT} ($n=6$) and *PIK3CA*$^{CAGG\text{-}CreER}$ ($n=6$) mice 7 days after they received a single dose of 40 mg kg^{-1} tamoxifen to induce *Cre* recombination. MRI scans were then repeated weekly in all mice for 1 month.

Flow cytometry. Spleens of *PIK3CA*^{WT} ($n=12$) mice, *PIK3CA*$^{CAGG\text{-}CreER}$ mice induced with 40 mg kg^{-1} tamoxifen ($n=6$) and *PIK3CA*$^{CAGG\text{-}CreER}$ mice induced with 4 mg kg^{-1} tamoxifen ($n=6$) were mechanically disrupted in PBS/SVF 2%. Following dissociation, spleens were filtered, centrifuged and resuspended. Cells were then treated with an FC blocker for 10 min at 4 °C (Biolegend, ref# 101302) and fixed/permeabilized (BD Bioscience, ref# 554714). Cells were labelled with chicken polyclonal anti-GFP antibody (Abcam, ref# ab13970) for 30 min at 4 °C. Subsequently, cells were incubated with Alexa Fluor 647-labelled goat anti-chicken IgY antibody (Abcam, ref# ab150171) for 30 min at 4 °C. A background control incubated only with secondary antibody was also included. Samples were analysed using Gallios Flow Cytometer (Beckman Coulter) and FlowJo software (TreeStar). The mosaicism was assessed by the ratio of the number of GFP-positive cells to the total number of cells.

Patients. The study was conducted on 19 patients (4 adults and 15 children) followed at Necker hospital. This protocol was approved by the 'Agence Nationale de Sécurité du Médicament et des Produits de Santé' (ANSM, authorizations n°: 553984–553986, 584018, 585881–586135, 585464, 585465, 585467, 585880, 586136, 585463, 596229, 588018, 587904, 587896–587912, 587908, 587905, 587910, 585458, 595374 and 587899) and informed written consent was obtained from the adult patients and the legal representatives of the children. BYL719 was compassionately offered by Novartis. Adult patients received 250 mg per day and child patients received 50 mg per day. BYL719 was taken orally every morning before breakfast.

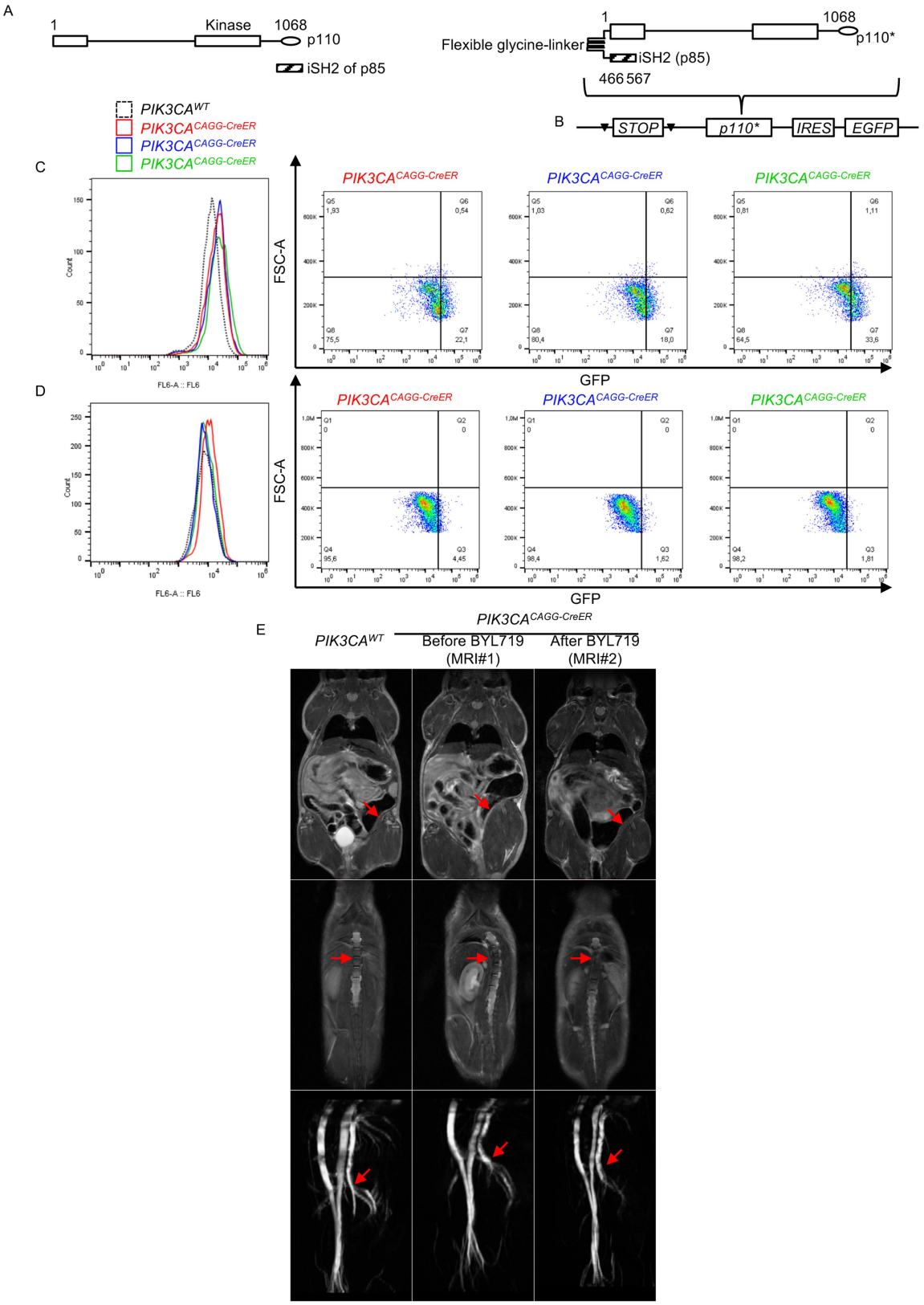
Patients were followed at regular intervals: weekly for eight weeks, every two weeks for one month and then monthly. Glycaemia was evaluated daily for the first month and then progressively less frequently. At all time points, the patients had a physical examination and performance status measurement using the Karnofsky (on a scale from 0 to 100, with lower numbers indicating greater disability) and ECOG indexes (a scale of 0 to 5, with 0 indicating no symptoms and higher scores indicating increasing symptoms)[23,24]. Growth of the children was monitored at all clinical appointments. Blood sampling (complete blood count, kidney and liver functions, glycated haemoglobin measurement) were performed at each time points. Glycaemia was monitored after all meals for two months and then the monitoring became progressively less frequent. Adverse events were graded according to National Cancer Institute Common Terminology Criteria for Adverse Events, version 4.0. All patients had heart ultrasound examination before BYL719 treatment and then every three months. Magnetic resonance imaging studies were performed before BYL719 introduction and then after three and six months of treatment. PET scans were performed before and at 3 months of treatment.

Data analysis and statistics. Data are expressed as mean ± s.e.m. Survival curves were analysed with the Mantel–Cox (log-rank) test. Differences between the experimental groups were evaluated using ANOVA, followed when significant ($P<0.05$) with the Tukey–Kramer test. When only two groups were compared, Mann–Whitney tests were used. Statistical analyses were performed using GraphPad Prism software.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

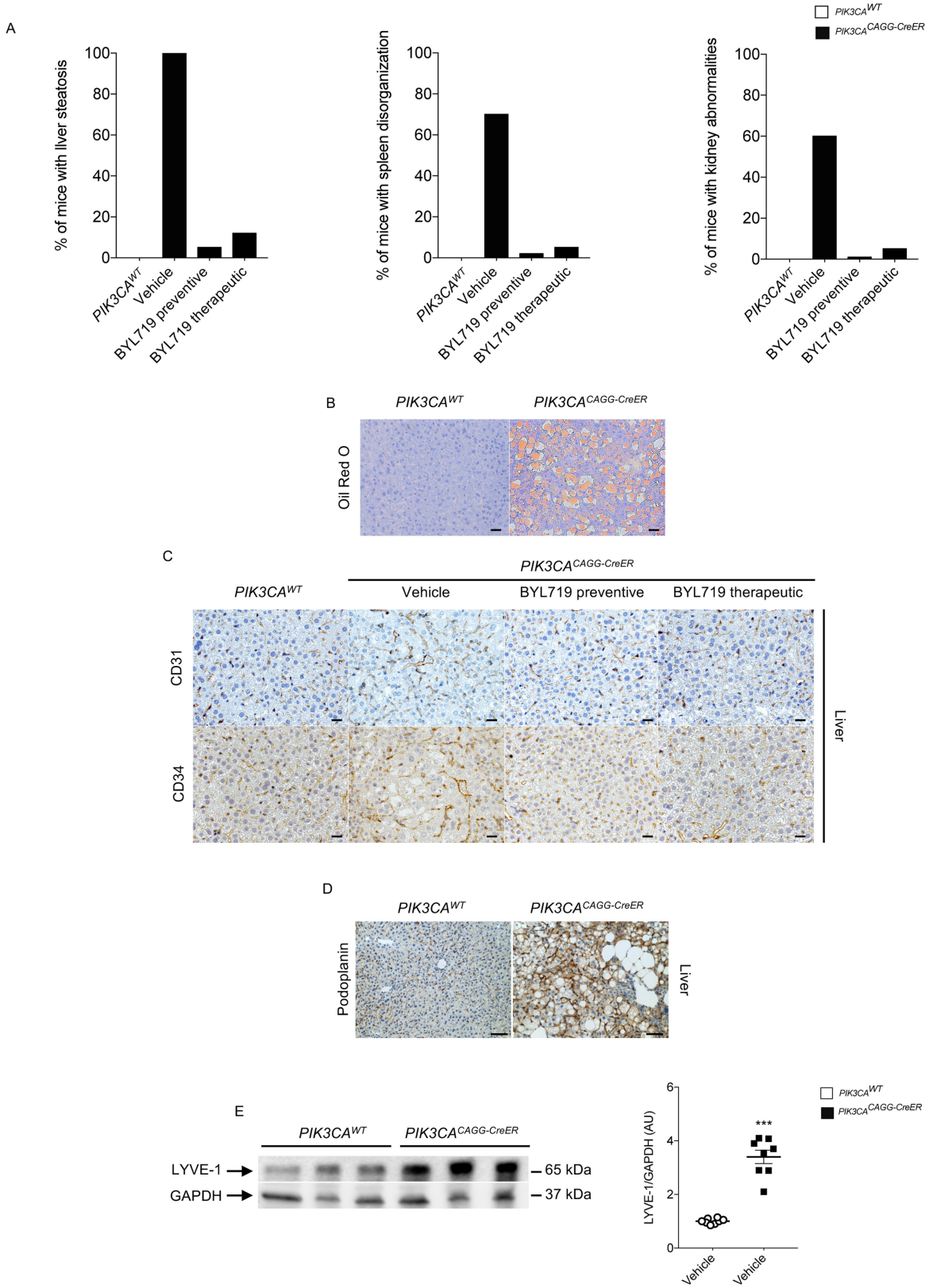
Data availability. All data generated or analysed during this study are included in this published article (and its Supplementary Information Files).

25. Seluanov, A., Vaidya, A. & Gorbunova, V. Establishing primary adult fibroblast cultures from rodents. *J. Vis. Exp.* **2010**, 2033 (2010).
26. Vandesompele, J. et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, RESEARCH0034 (2002).
27. Canaud, G. et al. Inhibition of the mTORC pathway in the antiphospholipid syndrome. *N. Engl. J. Med.* **371**, 303–312 (2014).

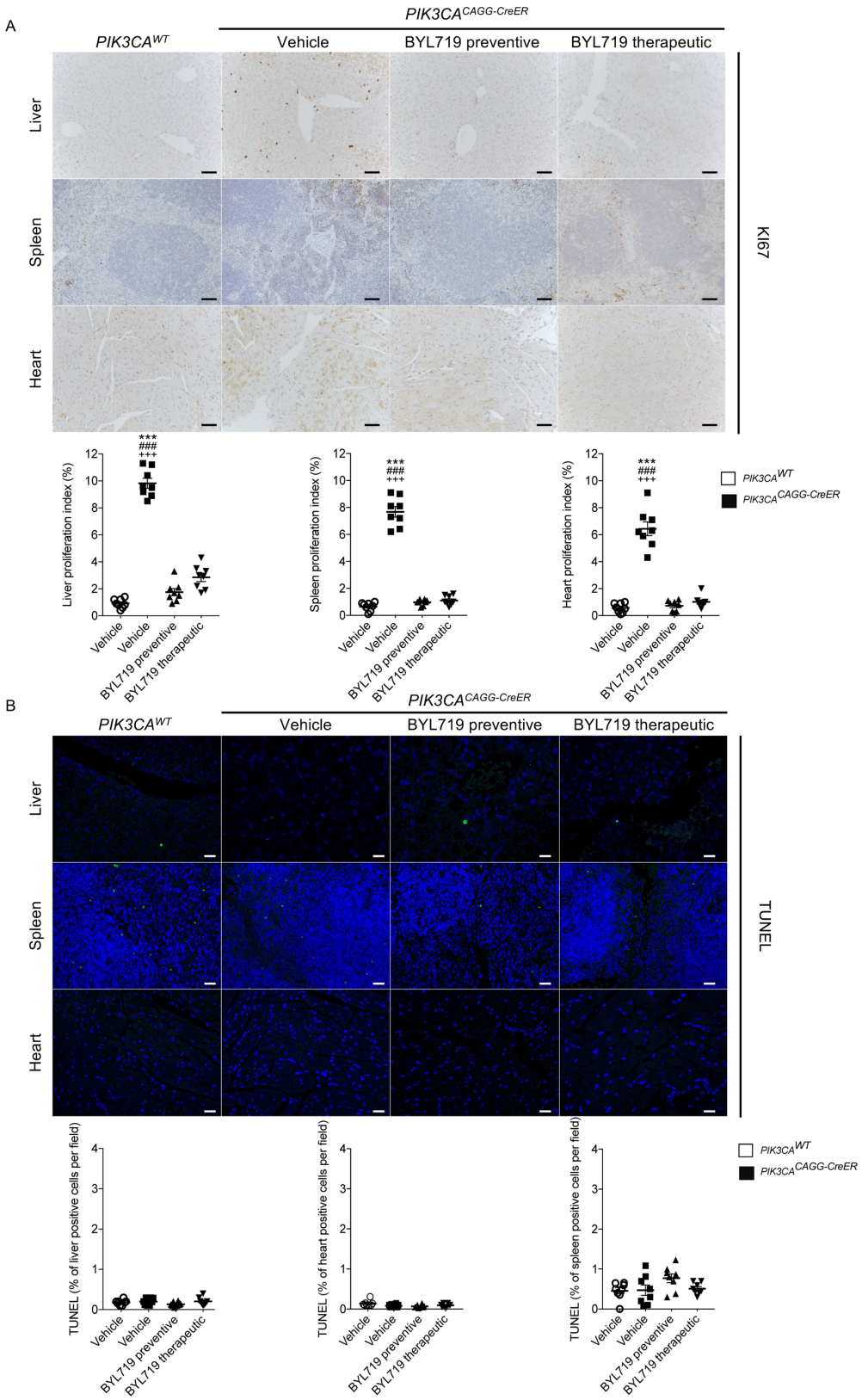


Extended Data Fig. 1 | See next page for caption.

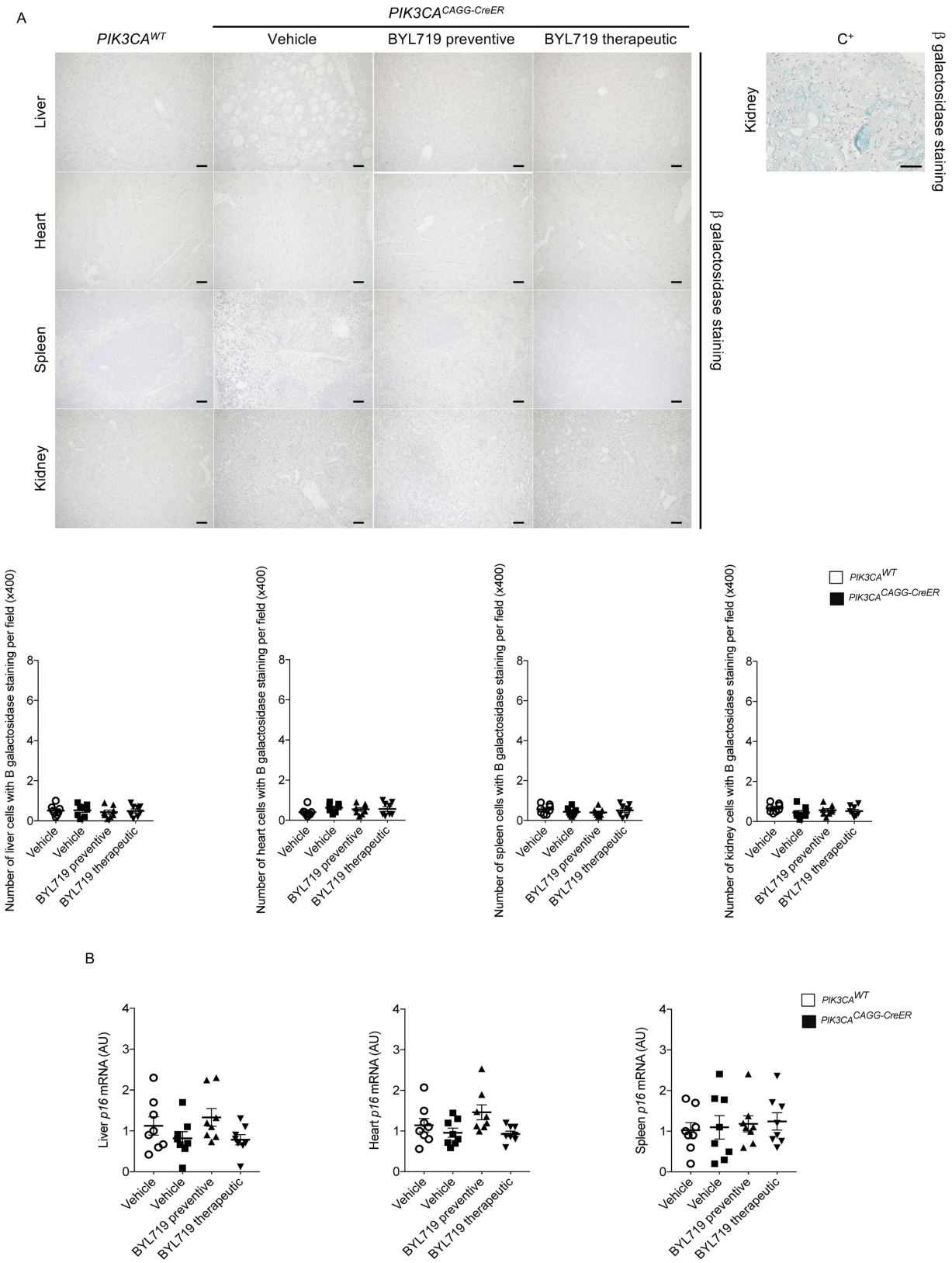
Extended Data Fig. 1 | p110* construction and mouse model characterization. a, Left, Representation of p110 and iSH2 domain of the p85 subunit (striped bar). The iSH2 domain is important to stabilize the p110α protein. The p110* protein is a constitutively active chimaera that contains the iSH2 domain of p85 fused to the N terminus of p110 via a flexible glycine linker[14] (right). **b**, To generate tissue-specific p110*-transgenic mice, a cloned loxP-flanked neoR-stop cassette was inserted into a modified version of pROSA26-1 followed by the cDNA encoding p110* and then a frt-flanked IRES–EGFP cassette and a bovine polyadenylation sequence (*R26StopFLP110**)[13]. **c**, **d**, EGFP expression from flow cytometry experiments in the spleen of $PIK3CA^{WT}$ mice ($n = 12$) and $PIK3CA^{CAGG\text{-}CreER}$ mice injected with either a single 40 mg kg^{-1} dose (**c**; $n = 6$ mice) or a single 4 mg kg^{-1} dose (**d**; $n = 6$ mice) of tamoxifen. Each curve is a different mouse. **e**, MRI examination of the PROS mouse model and efficacy of BYL719 treatment. Top, arrows show muscle hypertrophy in $PIK3CA^{CAGG\text{-}CreER}$ mice before BYL719 treatment. This phenotype was reversed by BYL719 administration. Middle, arrows show scoliosis in $PIK3CA^{CAGG\text{-}CreER}$ mice before BYL719 treatment, which was rescued by BYL719 administration. Bottom, arrows show arterial dilation in $PIK3CA^{CAGG\text{-}CreER}$ mice before BYL719 treatment, which was reversed by BYL719 administration ($n = 6$ mice per group).



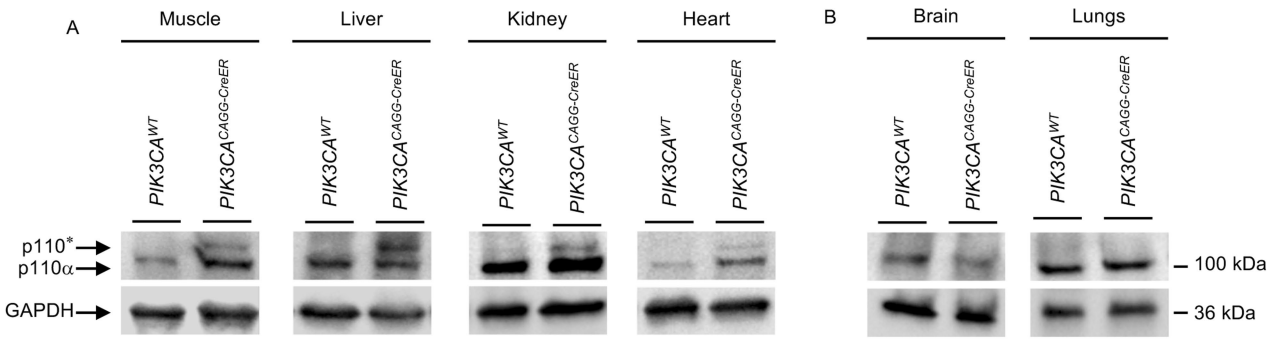
Extended Data Fig. 2 | Quantification and vessel malformation. **a**, Percentage of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice with or without BYL719 treatment presenting organ abnormalities. **b**, Oil Red O staining of the livers of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice demonstrating steatosis ($n = 8$ mice per group). Scale bars, 10 μm. **c**, CD31 (top) and CD34 (bottom) immunostaining in the liver of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice with or without BYL719 ($n = 8$ mice per group). *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with vehicle showed vessel dilation that was prevented or reversed by BYL719. Scale bars, 10 μm. **d**, Representative picture of lymphatic malformation as assessed by podoplanin immunostaining in the liver of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice ($n = 8$ mice per group). Scale bars, 10 μm. **e**, Representative western blot of LYVE-1 in the liver of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice demonstrating lymphatic increased in the *PIK3CA*$^{CAGG\text{-}CreER}$ mice ($n = 8$ mice per group). All data are shown as the means ± s.e.m. Mann–Whitney test (two-tailed, $P = 0.001$). *PIK3CA*$^{CAGG\text{-}CreER}$ versus *PIK3CA*^{WT} mice, $^{***}P < 0.001$.

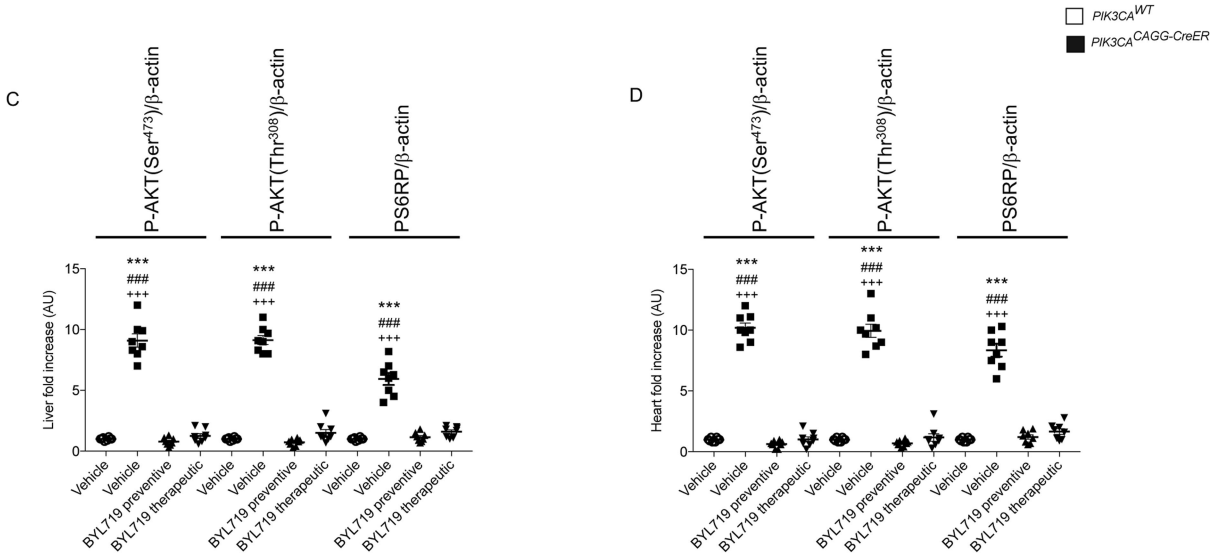


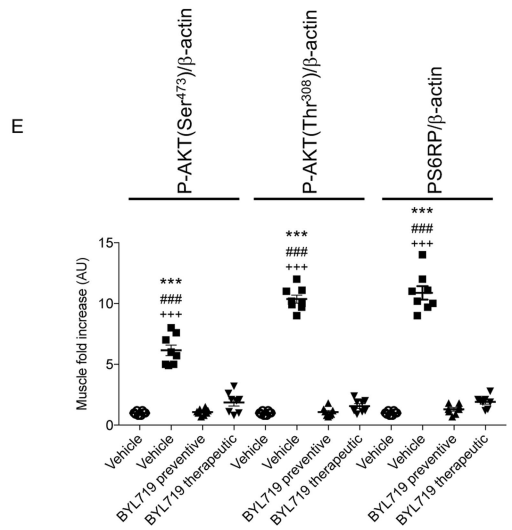
Extended Data Fig. 3 | BYL719 affects proliferation. a, Ki67 immunostaining and quantification in liver, spleen and heart of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice with or without BYL719 treatment ($n = 8$ mice per group, 10 randomly selected fields per mice, ×400). **b**, TUNEL assay. The graphs show the quantification of TUNEL-positive cells per field ($n = 8$ mice per group, 10 randomly selected fields per mice, ×400). Scale bars, 10 μm. All data are shown as mean ± s.e.m. ANOVA followed by Tukey–Kramer test (two-tailed). *PIK3CA*$^{CAGG\text{-}CreER}$ versus *PIK3CA*^{WT} mice, $^{***}P < 0.001$. *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with vehicle versus *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with preventive BYL719, $^{\#\#\#}P < 0.001$. *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with vehicle versus *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with therapeutic BYL719, $^{+++}P < 0.001$.



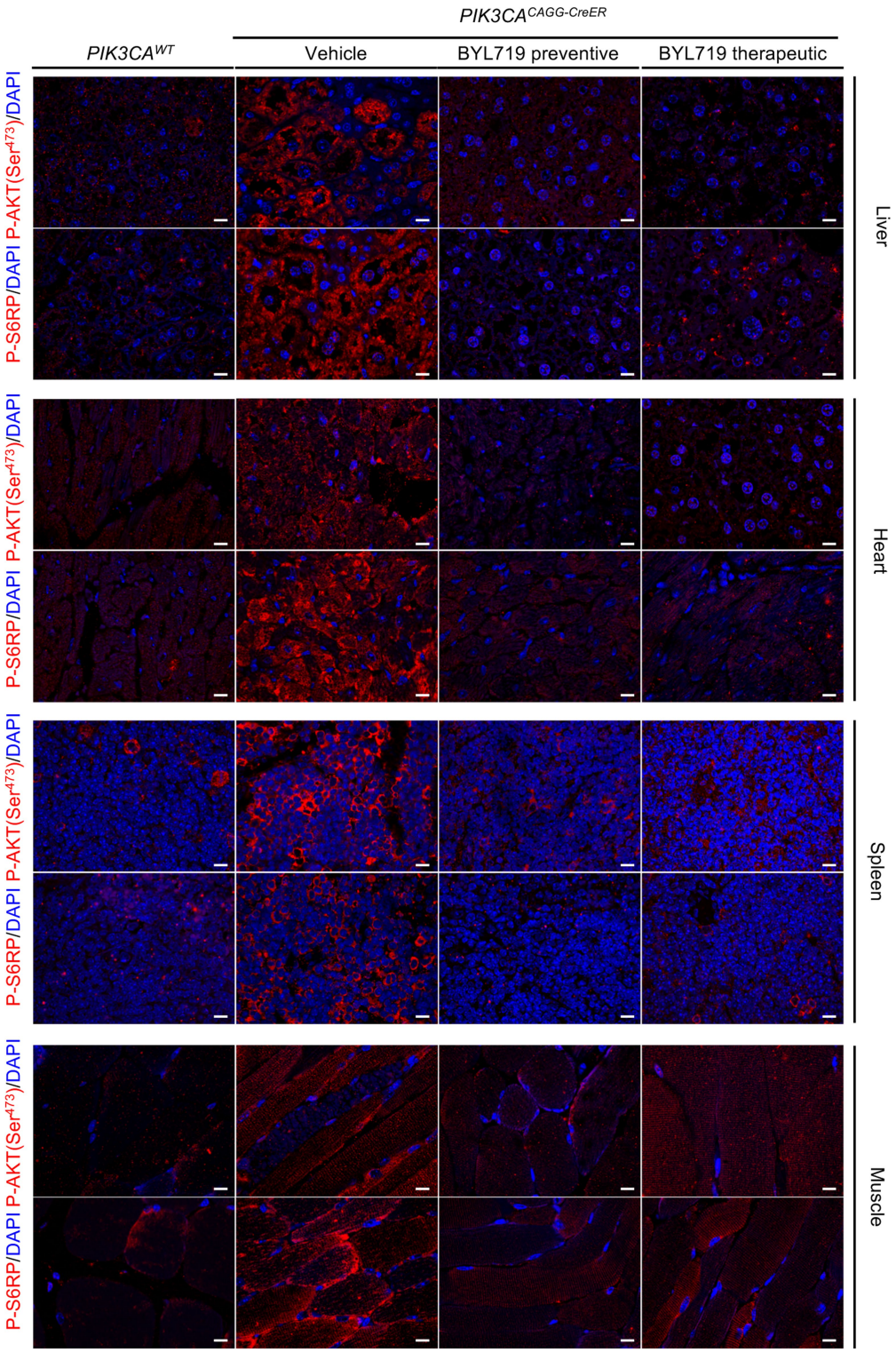
Extended Data Fig. 4 | Senescence and BYL719. a, β-galactosidase staining in the liver, heart, spleen and kidney of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice with or without BYL719 and quantification of β-galactosidase-positive cells per field ($n = 8$ mice per group, 10 randomly selected fields, ×400). C^{+}: positive control. Scale bars, 10 μm. **b**, *p16* mRNA expression in liver, heart and spleen of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with or without BYL719 ($n = 8$ mice per group). A.U., arbitrary unit. All data are shown as mean ± s.e.m. ANOVA followed by Tukey–Kramer test (two-tailed).



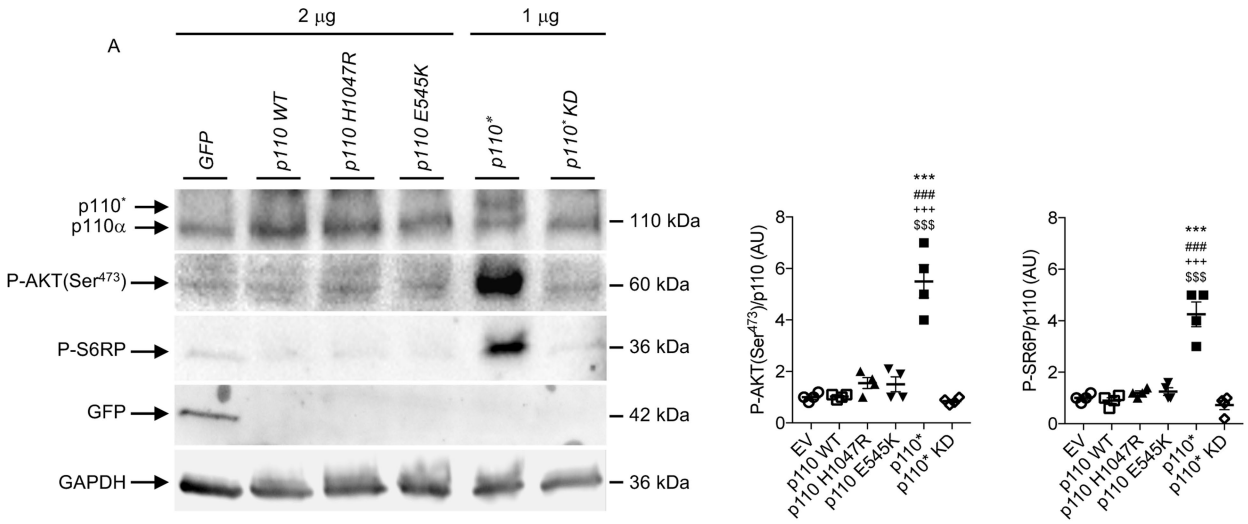


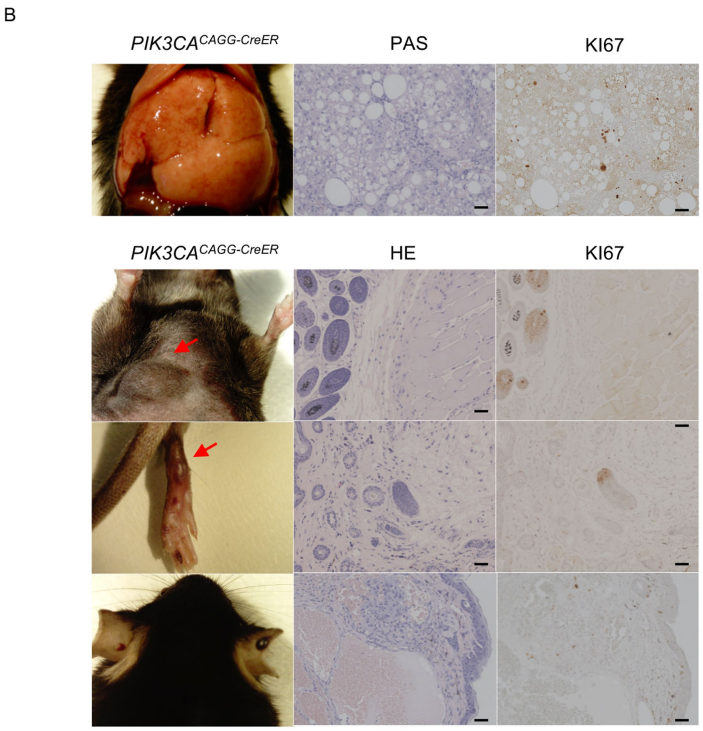


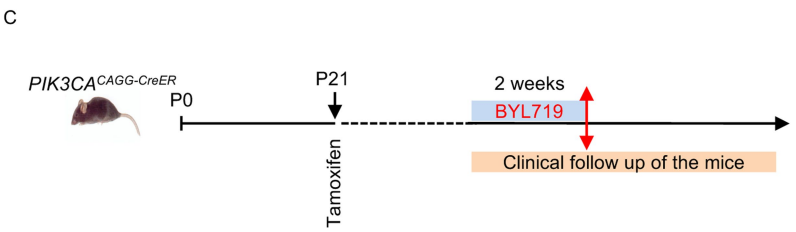
Extended Data Fig. 5 | p110* expression in affected tissues. a, Western blot showing the expression of p110* in *PIK3CA*$^{CAGG\text{-}CreER}$ mice ($n = 8$ mice per group). **b**, p110* is not expressed in the brain or lungs ($n = 8$ mice per group). **c**–**e**, Western blot quantification of Fig. 1d, in the liver (**c**), heart (**d**) and muscle (**e**) of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with or without BYL719 ($n = 8$ mice per group). All data are shown as mean ± s.e.m. ANOVA followed by Tukey–Kramer test (two-tailed). *PIK3CA*$^{CAGG\text{-}CreER}$ versus *PIK3CA*^{WT} mice, $^{***}P < 0.001$. *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with vehicle versus *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with preventive BYL719, $^{\#\#\#}P < 0.001$. *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with vehicle versusd *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with therapeutic BYL719, $^{+++}P < 0.001$.



Extended Data Fig. 6 | Ability of BYL719 to inhibit PIK3CA activation in different tissues. Immunofluorescence staining of P-AKT (Ser^{473}) and P-S6RP in the liver (**a**), heart (**b**), spleen (**c**) and muscles (**d**) of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with or without BYL719 ($n = 8$ mice per group). Scale bars, 10 μm.

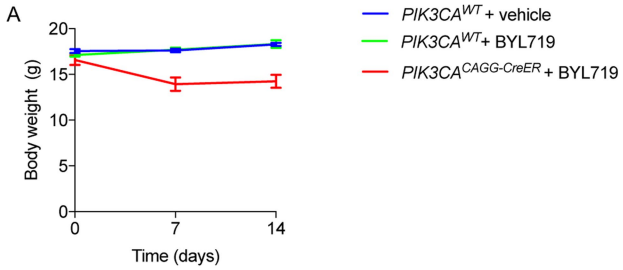


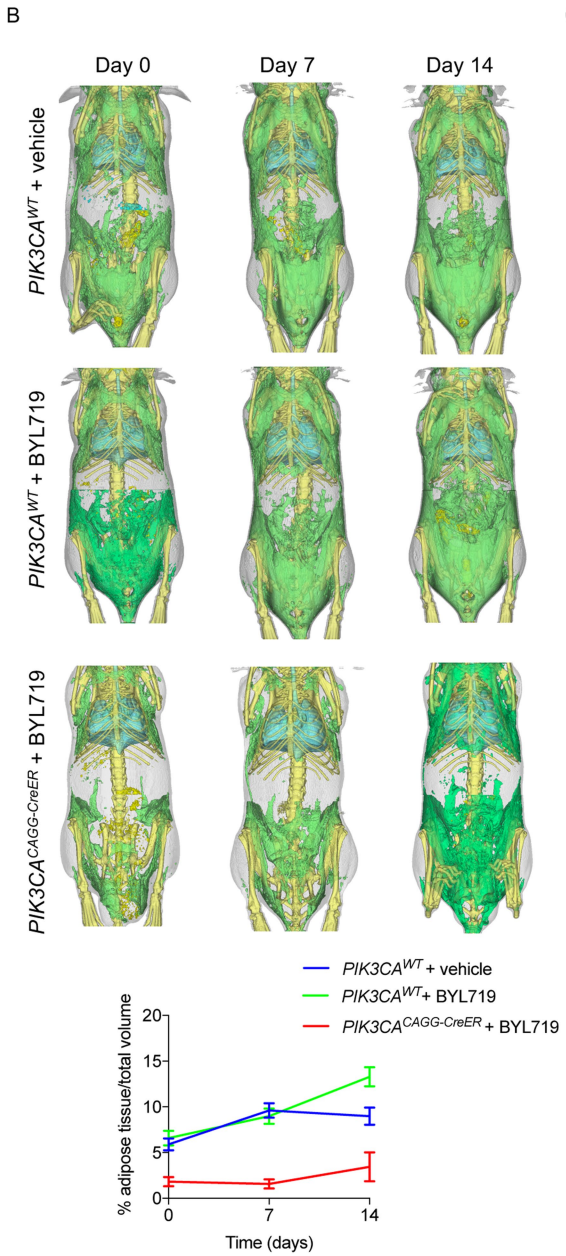


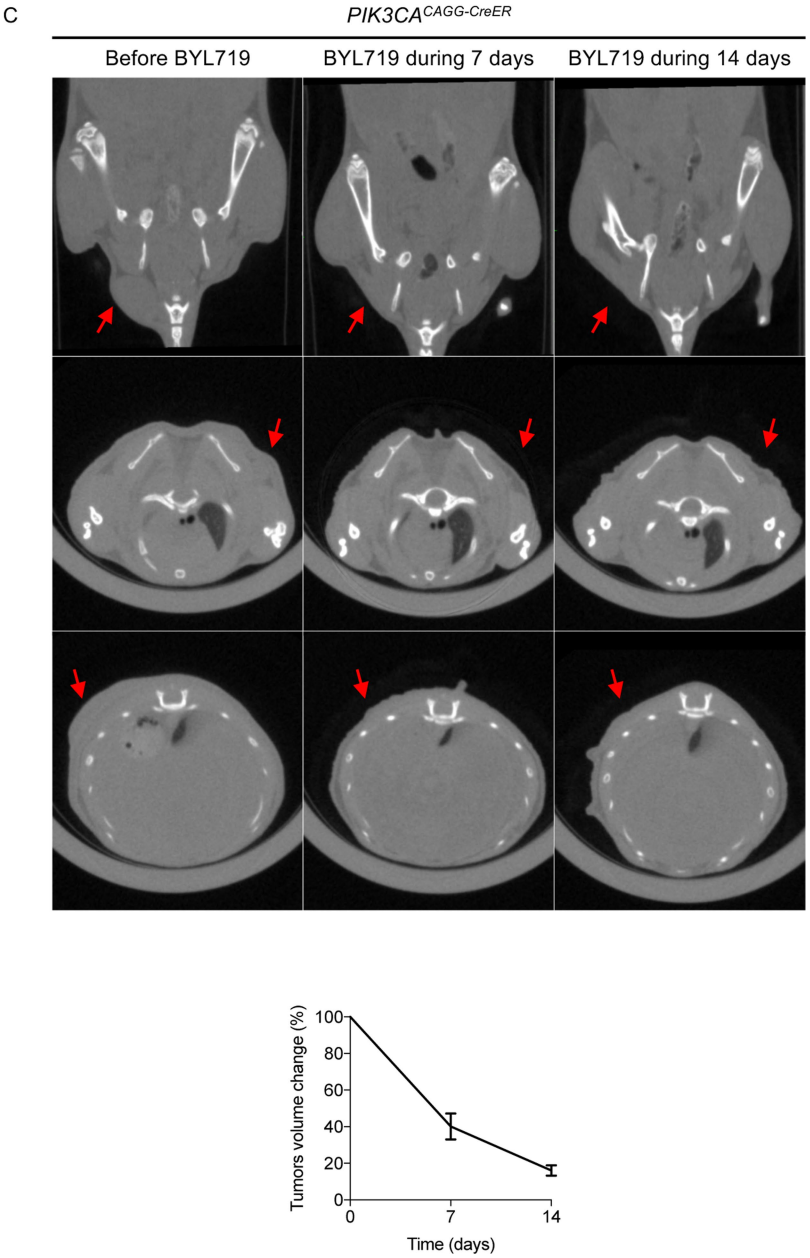


Extended Data Fig. 7 | See next page for caption.

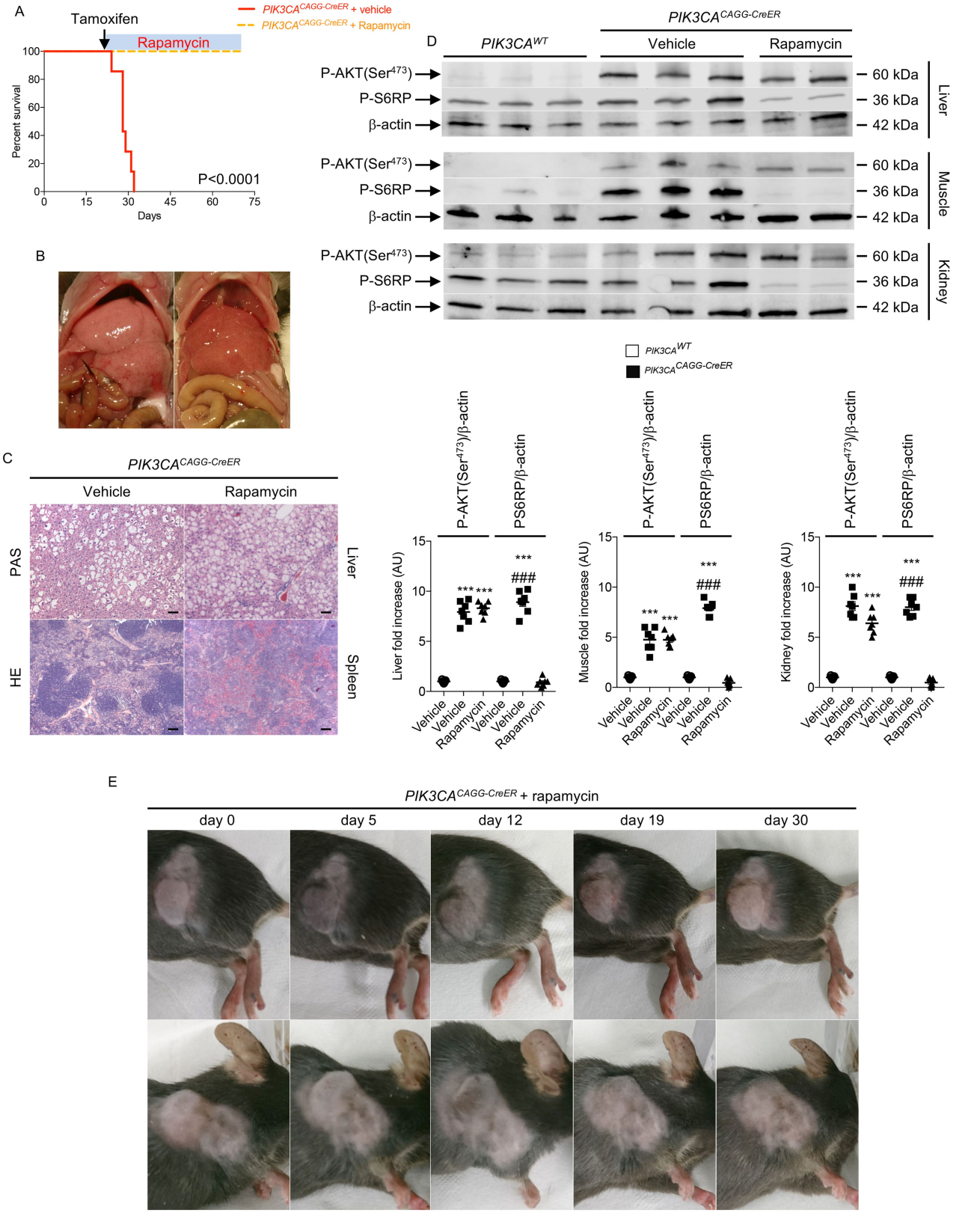
Extended Data Fig. 7 | Recruitment of the AKT/mTORC pathway by the different forms of mutant p110. a, Western blot and quantification of p110, P-AKT (Ser^{473}), P-S6RP and GFP in HeLa cells transfected with plasmids containing cDNA encoding p110*, p110* kinase-dead mutant (p110* KD) as a control, H1047R mutation or E545K mutation. Cells transfected with the p110* mutant showed a more powerful effect on the activation of the AKT/mTORC pathway than the others ($n = 4$ independent experiments). All data are shown as mean ± s.e.m. ANOVA followed by Tukey–Kramer test (two-tailed). p110* versus H1047R mutation, $^{***}P < 0.001$. p110* versus E545K mutation, $^{\#\#\#}P < 0.001$. p110* versus wild-type p110, $^{+++}P < 0.001$. p110* versus p110* KD, $^{\$\$\$}P < 0.001$. Negative control is a vector that contains cDNA encoding GFP. **b**, Histological examination of different tissues from $PIK3CA^{CAGG\text{-}CreER}$ mice. Left column, from top to bottom, liver, abdominal tumour, leg and ear abnormalities. Middle column, PAS or HE staining of the tissue. Right column, Ki67 staining of the same tissue ($n = 8$ mice). Scale bars, 20 μm. **c**, Design of the experiment shown in Fig. 2h. $PIK3CA^{CAGG\text{-}CreER}$ mice received a single dose of 4 mg kg^{-1} tamoxifen and were followed for one month. Once the tumours became visible, BYL719 was started for two weeks and then withdrawn.



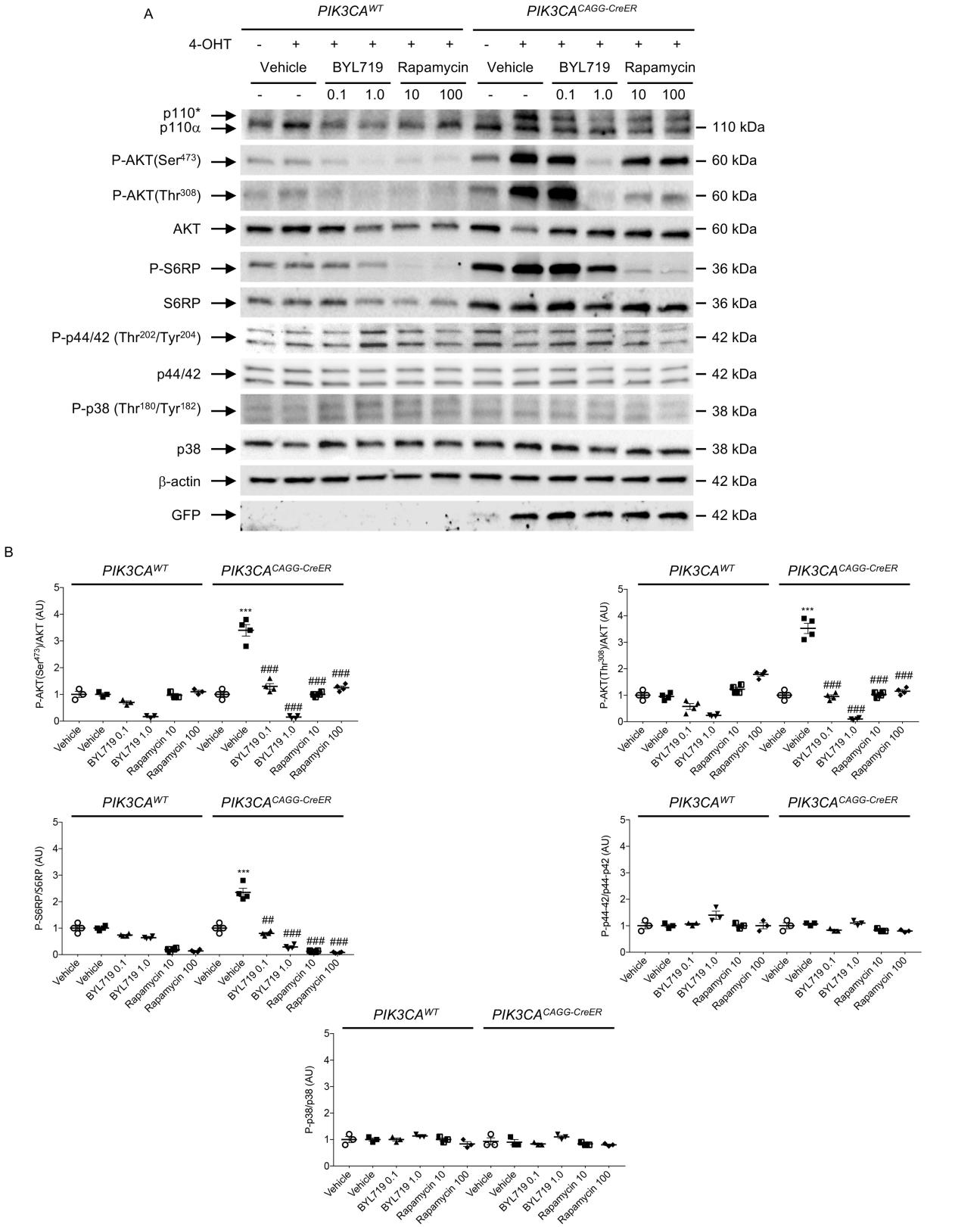




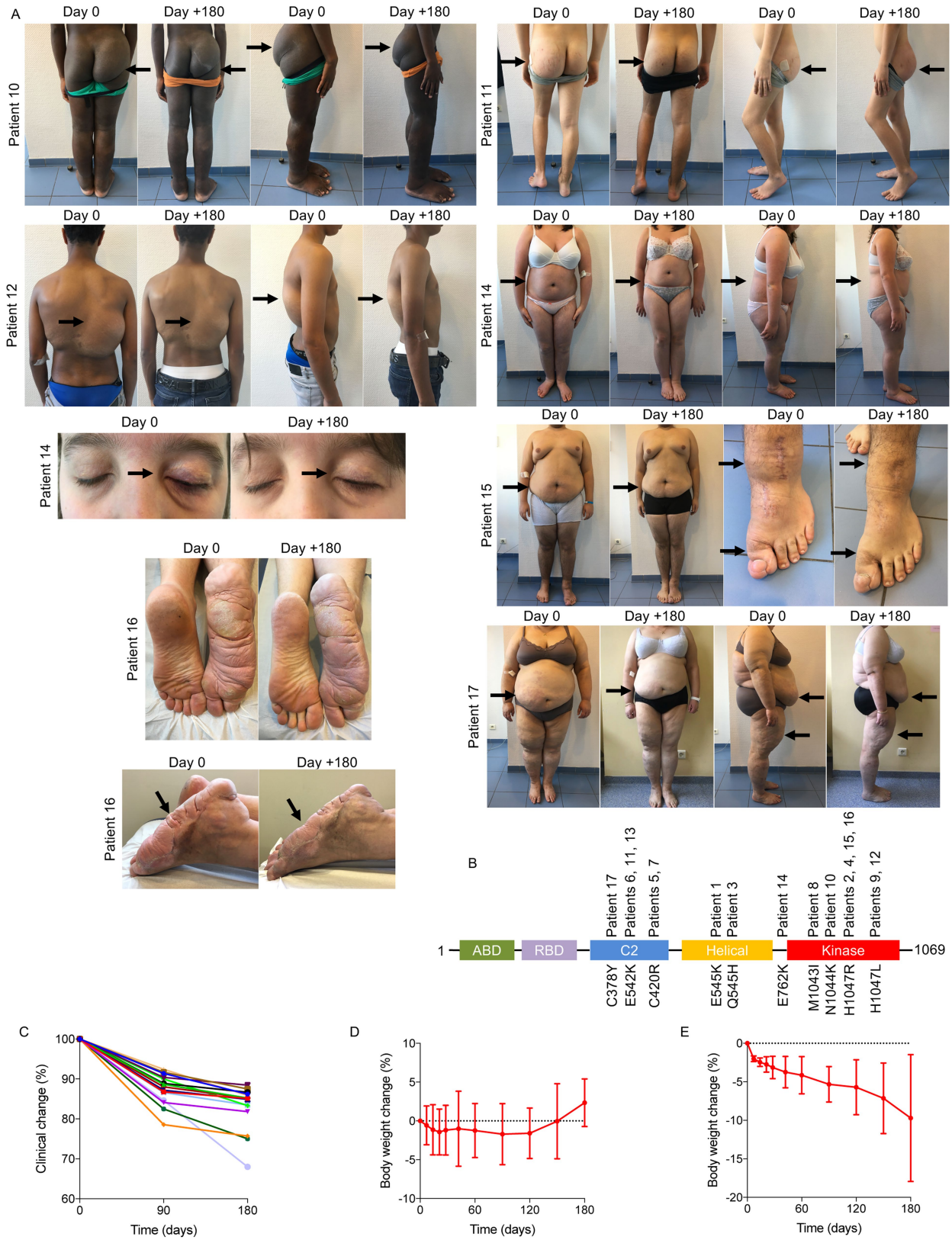
Extended Data Fig. 8 | CT scan evaluation of the tumours and adipose tissue before and after BYL719 introduction. a, Body weight evolution of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with vehicle or BYL719 ($n = 3$ mice per group). **b**, CT scan evaluation and quantification of the fat tissue content in *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with vehicle or BYL719. Subcutaneous and visceral fat content were measured before treatment and 7 and 14 days after onset of treatment with vehicle or BYL719 ($n = 3$ mice per group). **c**, CT scan evaluation and quantification of tumour volume in *PIK3CA*$^{CAGG\text{-}CreER}$ mice before and after two weeks of BYL719 treatment (arrows) ($n = 3$ mice per group). All data are shown as mean ± s.e.m.



Extended Data Fig. 9 | Effect of rapamycin treatment on the different *PIK3CA*$^{CAGG\text{-}CreER}$ mouse models. a, Kaplan–Meier survival curves of *PIK3CA*$^{CAGG\text{-}CreER}$ mice that received a single dose of 40 mg kg^{-1} tamoxifen and were treated with or without rapamycin after tamoxifen administration. **b**, Representative pictures of the liver of *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with rapamycin 40 days after *Cre* induction. **c**, Morphology of livers and spleens from *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice that were treated with or without rapamycin after *Cre* induction. Scale bars, 10 μm. **d**, Western blot and quantification of P-AKT (Ser^{473}) and P-S6RP in the liver, heart and muscle, respectively, of *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with vehicle or rapamycin directly after *Cre* induction. **e**, *PIK3CA*$^{CAGG\text{-}CreER}$ mice were treated with rapamycin one month after *Cre* induction with a single dose of 4 mg kg^{-1} tamoxifen and followed for one month. All data are shown as mean ± s.e.m. ANOVA followed by Tukey–Kramer test (two-tailed). *PIK3CA*$^{CAGG\text{-}CreER}$ versus *PIK3CA*^{WT} mice, $^{***}P < 0.001$. *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with rapamycin compared with *PIK3CA*$^{CAGG\text{-}CreER}$ mice treated with vehicle, $^{\#\#\#}P < 0.001$.

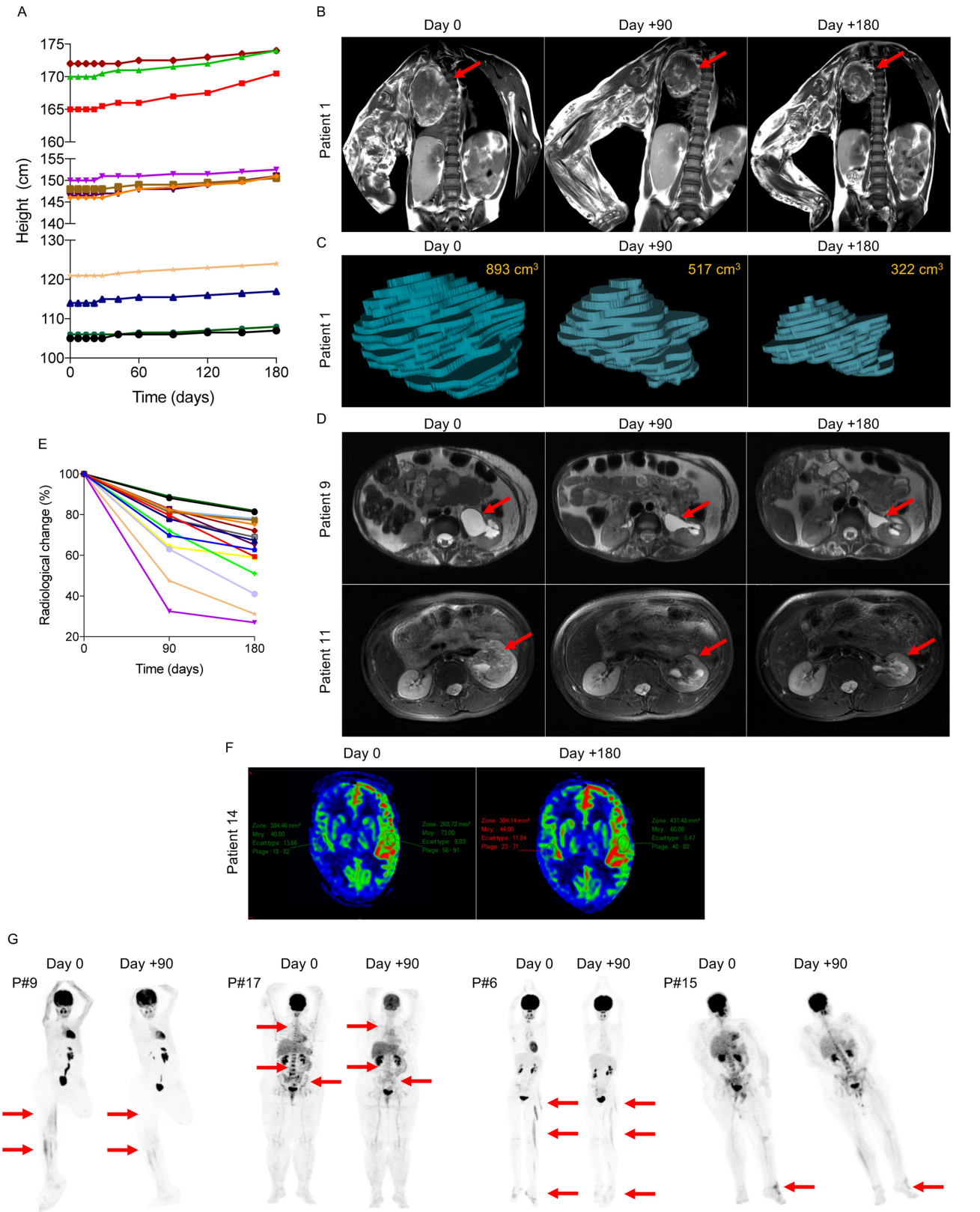


Extended Data Fig. 10 | In vitro effect of BYL719 and rapamycin on fibroblasts from *PIK3CA*$^{CAGG\text{-}CreER}$ mice. a, Skin fibroblasts from *PIK3CA*^{WT} and *PIK3CA*$^{CAGG\text{-}CreER}$ mice were isolated and exposed to vehicle or increasing concentrations of BYL719 or rapamycin for 24 h. **b**, Quantification. White column, without 4-OHT; black column, with 4-OHT. All data are shown as mean ± s.e.m. ANOVA followed by Tukey–Kramer test (two-tailed). Before versus after *Cre* induction with 4-OHT, $^{***}P < 0.001$. BYL719 or rapamycin exposure compared with cells treated with vehicle, $^{\#\#}P < 0.01$ and $^{\#\#\#}P < 0.001$.



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Effect of BYL719 in patients with PROS. **a**, Patients 10–17 before and after 180 days of BYL719 treatment. Patient 10 was a 14-year-old boy with severe asthenia, dyspnea and bilateral overgrowth of lower limbs. After 180 days of treatment asthenia resolved and we observed a marked reduction in hypertrophy of the limbs. Patient 11 was a 14-year-old boy with overgrowth of the right buttock and an intra-abdominal vascular tumour infiltrating the left kidney and spinal nerve. He had chronic haematuria and was permanently confined to bed owing to pain. After 180 days, haematuria resolved and the volume of the intraabdominal vascular malformation was reduced by up to 68%. He had no more pain and became capable of walking. Patient 12 was a 15-year-old boy with multiple large tumours of the trunk and the back. After 180 days of treatment the tumours had reduced in size. Patient 13 was a 16-year-old boy with megalencephaly-capillary malformation (MCAP) and left hemifacial hyperplasia. Treatment led to a reduction in hemifacial hypertrophy and cognitive improvement. Owing to the deformation, this patient was not able to open the left eye. After 180 days of BYL719 treatment, he was able to open the eye (not shown for confidentiality reasons). Patient 14 was a 16-year-old girl with MCAP and a chronic noninfectious palpebral cellulitis who was steroid-dependent. BYL719 treatment led to the healing of the cellulitis and steroids were stopped without a flare. We also observed enhancement of cognitive function and behaviour and improvement of scoliosis. Patient 15 was a 19-year-old man with overgrowth of the left foot and unstable and painful walking. Treatment led to an improvement in the overgrowth as well as an improvement in walking distance. Patient 16 was a 32-year-old man with overgrowth of the right foot and unstable and painful walking. Treatment led to an improvement in the overgrowth as well as an improvement in walking distance. Patient 17 was 50-year-old woman with generalized hypertrophy, and severe and diffuse pain with opioid dependency. She was permanently confined to bed. After six months of treatment we observed an improvement in tiredness, and resolution of pain, and we were able to stop opioids within two weeks. The patient became able to walk again. **b**, *PIK3CA* mutations identified in the 17 patients. **c**, For each patient we determined a target lesion (see Supplementary Table 2) that was clinically measured at each time point. The graph represents the changes (per cent) during the 180 days of treatment with BYL719. Each line is a single patient. **d**, Mean body weight changes (per cent) during the 180 days of treatment with BYL719 ($n = 13$ patients, patients 1–13), excluding the four obese patients (patients 14, 15, 16 and 17). **e**, Mean body weight loss in the four obese patients during the 180 days of treatment with BYL719. All data are shown as mean ± s.e.m.



Extended Data Fig. 12 | Height changes in children during treatment period and radiological changes with BYL719 treatment. a, Height changes in children during the 180 days of treatment with BYL719. **b**, MRI scans of patient 1 before and after 180 days of BYL719 treatment. Arrows show the target lesion. **c**, Three-dimensional MRI-based reconstruction of the chest tumour in patient 1 before and after 180 days of BYL719 treatment. **d**, Examples of MRI showing the evolution of the target lesions in patients 9 and 11. **e**, Volume evolution of the radiological target lesion after 180 days of BYL719 treatment. **f**, Diffusion MRI demonstrating the enhancement of brain perfusion in patient 14 after 180 days of BYL719. **g**, PET scan images of patients 6, 9, 15 and 17, before and after 90 days of BYL719 treatment. The arrows delineate hypermetabolic activity before and after 90 days of treatment.

ARTICLE

https://doi.org/10.1038/s41586-018-0219-7

Structure of the μ-opioid receptor–G_i protein complex

Antoine Koehl[1,12], Hongli Hu[1,2,12], Shoji Maeda[2,12], Yan Zhang[1,2], Qianhui Qu[1,2], Joseph M. Paggi[1,2,3,4], Naomi R. Latorraca[1,2,3,4,5], Daniel Hilger[2], Roger Dawson[6], Hugues Matile[6], Gebhard F. X. Schertler[7,8], Sebastien Granier[9], William I. Weis[1,2], Ron O. Dror[1,2,3,4,5], Aashish Manglik[10,11]*, Georgios Skiniotis[1,2]* & Brian K. Kobilka[2]*

The μ-opioid receptor (μOR) is a G-protein-coupled receptor (GPCR) and the target of most clinically and recreationally used opioids. The induced positive effects of analgesia and euphoria are mediated by μOR signalling through the adenylyl cyclase-inhibiting heterotrimeric G protein G_i. Here we present the 3.5 Å resolution cryo-electron microscopy structure of the μOR bound to the agonist peptide DAMGO and nucleotide-free G_i. DAMGO occupies the morphinan ligand pocket, with its N terminus interacting with conserved receptor residues and its C terminus engaging regions important for opioid-ligand selectivity. Comparison of the μOR–G_i complex to previously determined structures of other GPCRs bound to the stimulatory G protein G_s reveals differences in the position of transmembrane receptor helix 6 and in the interactions between the G protein α-subunit and the receptor core. Together, these results shed light on the structural features that contribute to the G_i protein-coupling specificity of the μOR.

The μOR is the primary target of morphine and many clinical opioid analgesics[1]. Binding of opioids to the μOR leads to clinically desired analgesic and antitussive actions, but also important negative side effects, including addiction and potentially lethal respiratory suppression. Opioids have become the most prescribed class of medication in the United States[2], which has led to a national epidemic of addiction and an unprecedented level of drug overdose deaths.

Like other GPCRs, the μOR achieves many of its physiological actions by stimulating signalling via a heterotrimeric G protein. While other GPCRs have been shown to signal through more than one G-protein subtype, the μOR signals almost exclusively through the adenylyl cyclase-inhibitory family of G proteins ($G_{i/o}$)[3]. The analgesic activity of opioids is driven by G-protein activation[4], but activated μOR can also interact with β-arrestins, recruitment of which has been associated with the respiratory depression induced by many opioids[5,6]. Recently developed molecules that favor G_i signalling over arrestin recruitment display analgesic efficacy with reduced side effects, suggesting that different signalling pathways can be selectively targeted to yield unique physiological outcomes[7,8]. Although a framework for GPCR interactions with the stimulatory G protein G_s has recently been enabled by X-ray crystallography[9] and cryo-electron microscopy (cryo-EM)[10,11] studies, the structural basis of GPCR signalling through other G-protein subtypes remains undefined. To better understand the mechanism of selective activation of G_i by the μOR, we sought to determine the structure of the μOR–G_i complex.

3.5 Å cryo-EM map of a μOR–G_i complex

DAMGO (H-Tyr-D-Ala-Gly-*N*(Me)Phe-Gly-OH) is a μOR-selective synthetic analogue of the natural peptide agonist enkephalin. DAMGO-bound μOR was incubated with G_{i1} heterotrimer, and the complex was treated with the nucleotide hydrolase apyrase to remove GDP. The resulting nucleotide-free complex was further stabilized by a single-chain variable fragment (scFv16) that binds to heterotrimeric G_i (Extended Data Fig. 1) and prevents GTPγS-mediated dissociation of nucleotide-free complexes. We applied single-particle cryo-EM to initially obtain a three-dimensional map of the μOR–DAMGO–G_i–scFv16 complex at an indicated nominal resolution of 3.6 Å (Extended Data Figs. 2, 3, Extended Data Table 1). Notably, scFv16 binds a composite interface comprised of the αN helix of $G\alpha_i$ and the β-propeller of Gβ, a site that is more than 20 Å distal to the μOR–$G\alpha_i$ interface and does not perturb the interface between Gα and Gβ subunits (Extended Data Fig. 1). Subtraction of the scFv16 signal from raw particle images led to an improved map with an indicated global resolution of 3.5 Å. This map displayed enhanced features, particularly in the transmembrane core of the receptor (Extended Data Figs. 2–4), which enabled the high-resolution visualization of the μOR–G_i interface and ligand binding. Accordingly, we employed this improved 3.5 Å map to examine interactions between μOR and DAMGO, and between μOR and G_i (Fig. 1a, b).

Activation of μOR by a peptide agonist

The active-state crystal structure of μOR bound to the morphinan agonist BU72 and an active-state stabilizing nanobody (Nb39) has been determined at a resolution of 2.2 Å[12]. Similar to other small molecule morphinans, BU72 is rigidified by a complex ring system; in contrast to flexible opioid peptides, such as DAMGO, that have multiple rotatable bonds. Our cryo-EM map includes well-defined features for most amino acids forming the orthosteric binding pocket (Extended Data Fig. 4a). Despite differences in agonist structure, the conformation of the active-state ligand-binding pocket and the orientation of the amino acids that interact with the agonist are highly similar between the μOR bound to either BU72 or DAMGO (Fig. 1c),

[1]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA. [2]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA. [3]Department of Computer Science, Stanford University, Stanford, CA, USA. [4]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. [5]Biophysics Program, Stanford University, Stanford, CA, USA. [6]Roche Pharma Research and Early Development, Therapeutic Modalities, Roche Innovation Center Basel, F.Hoffmann–La Roche, Basel, Switzerland. [7]Laboratory of Biomolecular Research, Paul Scherrer Institute, Villigen, Switzerland. [8]Department of Biology, ETH Zürich, Zürich, Switzerland. [9]Institut de Génomique Fonctionnelle, INSERM, Montpellier, France. [10]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, USA. [11]Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, CA, USA. [12]These authors contributed equally: Antoine Koehl, Hongli Hu, Shoji Maeda. *e-mail: Aashish.Manglik@ucsf.edu; yiorgo@stanford.edu; kobilka@stanford.edu

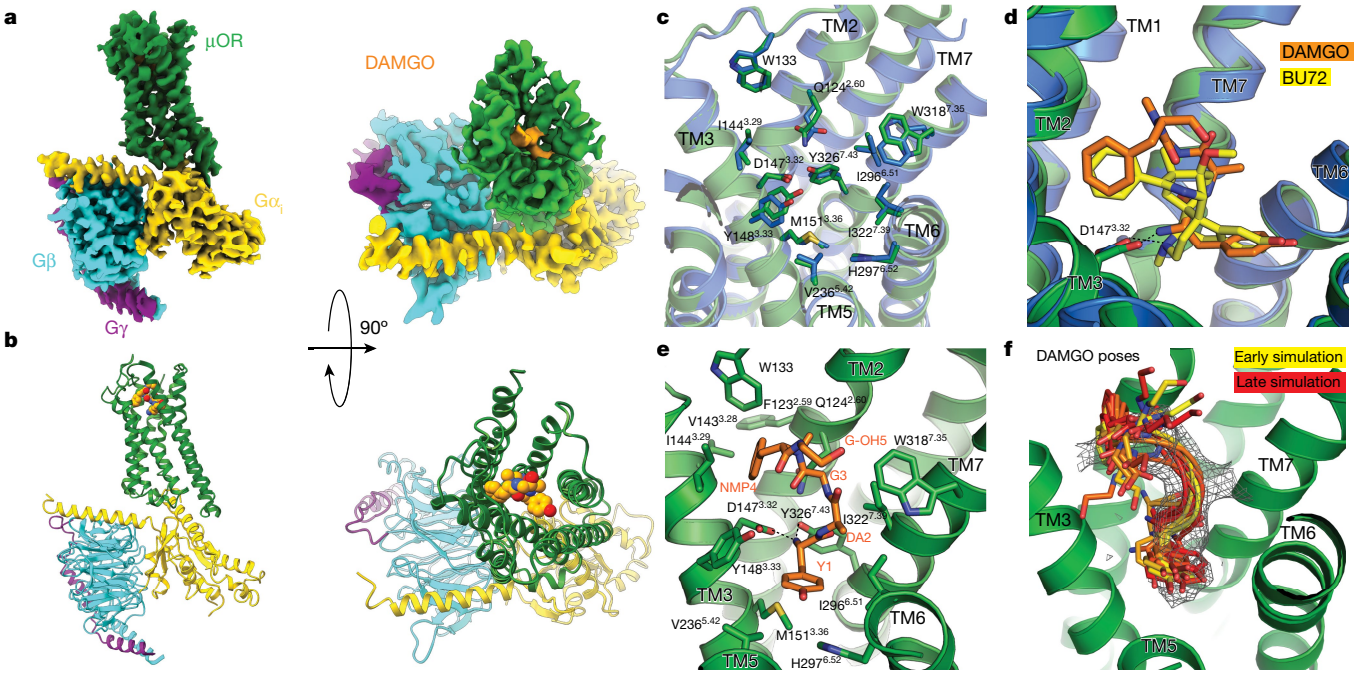


Fig. 1 | Cryo-EM structure of the μOR–G_i complex. a, Orthogonal views of the cryo-EM density map of the μOR–G_i heterotrimer complex coloured by subunit. Green, μOR; orange, DAMGO; gold, $G\alpha_i$ Ras-like domain; cyan, Gβ; purple, Gγ. **b**, Model of the μOR–G_i complex in the same views and colour scheme as shown in **a**. **c**, Residues that line the μOR orthosteric binding pocket are shown as sticks for the μOR–G_i complex (green) and the μOR–Nb39 complex (PDB code 5C1M; blue). The binding pocket residues of μOR in complex with DAMGO and BU72 show nearly identical conformations, despite differences in ligand structure. **d**, Comparison of BU72 (yellow carbons) in the orthosteric pocket of the μOR–Nb39 complex (blue) with DAMGO (orange carbons) in the orthosteric pocket of the μOR–G_i complex (green). **e**, view of DAMGO in the orthosteric binding pocket with critical residues shown. **f**, A frame from every 100 ns of a 1 μs molecular dynamics simulation (yellow at t = 0, fading to red at t = 1 μs) shows that the first four residues of DAMGO (bottom) are stable, whereas the C-terminal Gly-ol (top) is dynamic but frequently returns to the modelled pose.

suggesting that the μOR recognizes structurally distinct agonists in a stereotypic manner.

Although DAMGO is a flexible ligand, we observe density for the entire peptide bound to the receptor (Fig. 1a, Extended Data Figs. 3, 4). The N terminus of DAMGO occupies a similar position in the binding pocket as BU72. By contrast, the C terminus of DAMGO extends ~8 Å further towards the extracellular loops compared to BU72 (Fig. 1d, e). To identify stable atomic-level interactions between DAMGO and the binding pocket, we performed molecular dynamics simulations. In over 1 μs of simulation, DAMGO remained close to its initially modelled pose, with the N-terminal portion largely remaining confined to the experimentally determined cryo-EM density (Fig. 1f, Extended Data Fig 5). The N terminus of DAMGO maintained a persistent salt bridge with $D147^{3.32}$, a feature previously observed in structures of morphinans bound to opioid receptors (Fig. 1e; superscripts indicate Ballesteros–Weinstein numbering for GPCRs[13]). The same amine group also frequently formed a hydrogen bond with $Y326^{7.43}$. More generally, the N-terminal Tyr of DAMGO overlaps the phenolic groups of other small-molecule opioids that have been characterized in complex with μOR or other opioid receptors by X-ray crystallography[14–17].

Molecular dynamics simulations also revealed a water-mediated hydrogen bonding network that closely overlaps with the water network observed in the high-resolution crystal structure of μOR[12] (Extended Data Fig. 6). In particular, the simulations revealed a stable, water-mediated interaction that is formed between the phenol of DAMGO and $H297^{6.52}$. Although the crystal structure of the μOR bound to BU72 shows two water molecules bridging the DAMGO phenol and $H297^{6.52}$, simulations of μOR bound to DAMGO and other phenolic ligands[8,12] suggest that one of these water molecules rapidly dissociates, and that a single water is required for stable ligand binding. This interaction is a hallmark of opioid recognition that has been observed for morphinans in complex with the μOR[12,14] as well as other small molecule and peptide-mimetic agonists of the homologous δ- and κ-opioid receptors (δOR and κOR)[15,16,18].

DAMGO exhibits more than 500-fold selectivity for the μOR over the δOR and κOR[19]. Structural studies have shown that interactions of ligands with the extracellular loops encode ligand subtype specificity among closely related opioid receptors[15]. Indeed, DAMGO selectivity for μOR over δOR has been shown to depend on residues in extracellular loop (ECL) 1, whereas selectivity over κOR results from differences in ECL3[20]. The map density for the C-terminal residues of DAMGO is slightly weaker than for the amino terminus, consistent with increased mobility of this region in simulations (Fig. 1f, Extended Data Fig 5). In our model, the *N*(Me)Phe side chain of DAMGO occupies a conserved hydrophobic pocket near ECL1 and the Gly-OH group folds back over the ligand (Fig. 1e). This model is consistent with the high affinity of μOR binding to cyclized enkephalins, which bridge the +2 and +5 positions of the peptide[21].

Structure of G_i–stabilized active μOR

The overall structure of G_i-bound μOR is similar to the active conformation of the BU72-bound μOR stabilized by Nb39[12] (root mean square deviation of 1 Å) with a predominant outward displacement of transmembrane helix (TM) 6 from the heptahelical bundle relative to the inactive state (Fig. 2a, b). A number of highly conserved residues in the GPCR family have been shown to be important for receptor activation, including the $D^{3.49}R^{3.50}Y^{3.51}$, $N^{7.49}P^{7.50}XXY^{7.53}$ and conserved core triad ($I^{3.40}$, $P^{5.50}$, $F^{6.44}$) motifs. The conformation of each of these regions in the μOR–G_i complex is virtually identical to the active state observed in the complex with Nb39 (Fig. 2c). The structural similarity of μOR between Nb39 and G_i-bound states indicate that these changes underlie ligand-mediated activation and are not specific to a particular intracellular binder. Indeed, Nb39 and G_i promote a similar increase in agonist affinity[12], supporting a common mechanism of allosteric communication between the intracellular G-protein-coupling domain and the ligand-binding pocket[12].

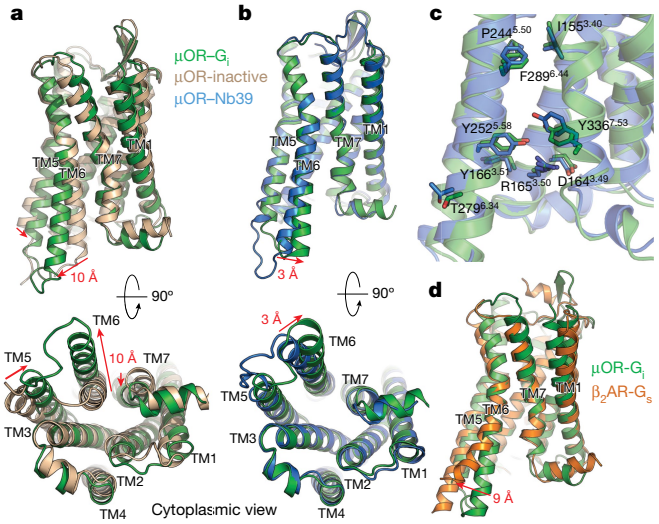


Fig. 2 | Structural changes in the μOR stabilized by nucleotide-free G_i. **a**, Comparison of inactive μOR (brown) and the G_i-stabilized active state of μOR (green). **b**, Comparison of Nb39-stabilized and G_i-stabilized active states of the μOR (blue and green, respectively). The structures are nearly identical except for a slight shift of TM6 towards TM7 in the G_i-bound state. **c**, Residues important for activation of the μOR show nearly identical conformations despite the difference in ligands. **d**, Comparison of G_s-stabilized β_2AR (orange) and G_i-stabilized μOR (green). While most transmembrane helices align well between the two receptors, TM6 is kinked further outward by 9 Å in the β_2AR; this distance is calculated between positions of C_α of residue 6.29 (Ballesteros–Weinstein numbering) in TM6.

Two differences between Nb39- and G_i-stabilized active states of μOR are particularly notable. First, compared with the nanobody-stabilized active-state μOR, TM6 in the μOR–G_i complex is further displaced by 3 Å towards TM7 (Fig. 2b). Second, the conformation of intracellular loop (ICL) 3 is different between the two structures (Fig. 2b). It is likely that the specific ICL3 conformation of Nb39-stabilized μOR reflects interactions that are unique to the nanobody rather than a general feature of receptor activation prior to G-protein coupling. A similar difference in ICL3 conformation was previously observed for the β_2-adrenergic receptor (β_2AR) between nanobody[22] (Nb80) and G_s-coupled states. The comparison of the G-protein-bound states of both receptors shows that TM6 of β_2AR is displaced outward by a further 9 Å in comparison to that of the μOR (Fig. 2d).

Structural changes in G_i

The quality of the cryo-EM map enabled accurate modelling of G_i in its nucleotide-free state, providing insight into the structural changes that underlie nucleotide release. The changes are similar to those seen in nucleotide-free G_s in complex with other GPCRs. The most striking difference between the GDP-bound[23] and nucleotide-free heterotrimer in complex with μOR involves the separation of the α-helical domain (AHD) from the Ras-like domain in $G\alpha_i$ (Fig. 3a). Owing to its relative flexibility, we excluded the AHD density from the high-resolution map refinement. The dynamic character of the AHD has been observed previously in spectroscopic and structural studies of complexes between receptors and G_s[9–11] and G_i[24,25]. Displacement of the AHD disrupts several contacts with GDP and is necessary, but not sufficient, for nucleotide release, a process that involves breaking additional contacts with the Ras domain[24].

Coupling of G_i to the μOR also involves a 6 Å translation as well as a 60° rotation of the α5-helix of $G\alpha_i$ into the receptor core (Fig. 3b). This movement has been shown to be essential for nucleotide release by G_i[24]. In particular, the movement of α5 leads to a change in the position of the β6–α5 loop containing the conserved TCAT motif that forms direct interactions with the guanine base of GDP. This displacement disrupts key contacts between the G protein and nucleotide. Furthermore, the observed translation and rotation of the α5-helix requires the displacement of the fully conserved F336 away from the hydrophobic pocket formed by residues in the β2 and β3 strands and the α1 helix[26] (Fig. 3b). Movement of the α5-helix is also propagated to the phosphate-binding P loop connecting the β1 strand and the α1 helix by disruption of a hydrophobic network between the α1 and α5 helices (Fig. 3b–d). Correspondingly, upon transition of G_i to the nucleotide-free state, we observe a 4 Å shift of α1 towards the α5-helix in G_i whereby the hydrophobic contacts are replaced by polar interactions with the β6–α5 loop as it is released from its guanine-binding position (Fig. 3c, d). These changes are in contrast to those observed in structures of G_s-coupled complexes, in which α1 not only becomes more unstructured, but also tends to lose interactions with the α5-helix (Fig. 3e). Our structure is consistent with previous studies suggesting that engagement of a GPCR with the α5-helix and αN–β1 loop leads to concerted changes in the α1 helix and P loop that destabilize contacts with the guanine nucleotide, leading to its release[27].

Structural insights into G_i–coupling specificity of the μOR

Although the μOR couples exclusively to $G_{i/o}$[3], many GPCRs can couple to multiple G-protein subtypes. A well-studied example is the β_2AR, which couples to both G_s and $G_{i/o}$. Previous sequence-level analyses have failed to identify a linear GPCR epitope that determines G-protein-coupling specificity, suggesting that specificity is likely to be determined by a more complex three-dimensional network of interactions. Globally, the structure of the μOR–G_i complex is similar to that of the β_2AR–G_s complex; this is likely to reflect a similarity in the conformation of nucleotide-free states of family A GPCR–G-protein complexes. The primary interaction sites in

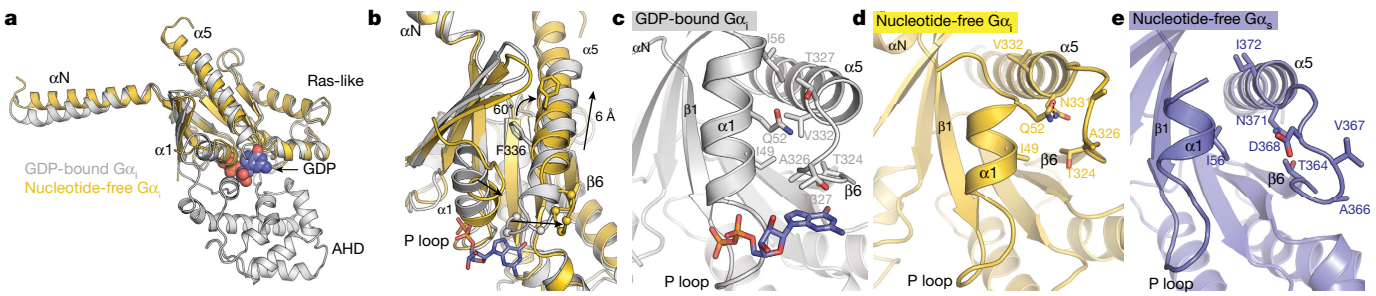


Fig. 3 | Changes in G_i upon coupling to the μOR. **a**, **b**, Comparison of GDP-bound $G\alpha_i$ (PDB code 1GP2, grey) and nucleotide-free $G\alpha_i$ from the μOR–G_i complex (gold). GDP is shown as blue spheres in **a** and sticks in **b** and **c**. The primary differences between these two structures are the opening and outward movement of the alpha helical domain (AHD), and an upward shift of the α5-helix by 6 Å to engage the receptor core. The α-carbons of the TCAT motif are represented as spheres in **b**. The TCAT motif coordinates the guanine base of GDP. The upward shift of the α5-helix and repositioning of the TCAT motif leads to nucleotide release. **c**, **d**, **e**, The interface between the α1 helix and the N-terminal end of the α5-helix and TCAT motif for GDP-bound $G\alpha_i$ (**c**), nucleotide-free $G\alpha_i$ (**d**) and nucleotide-free G_s from the β_2AR–G_s complex (**e**). The upward movement of the α5-helix disrupts the interaction between the α1- and α5-helices, leading to changes in the P loop that coordinates the phosphates of GDP.

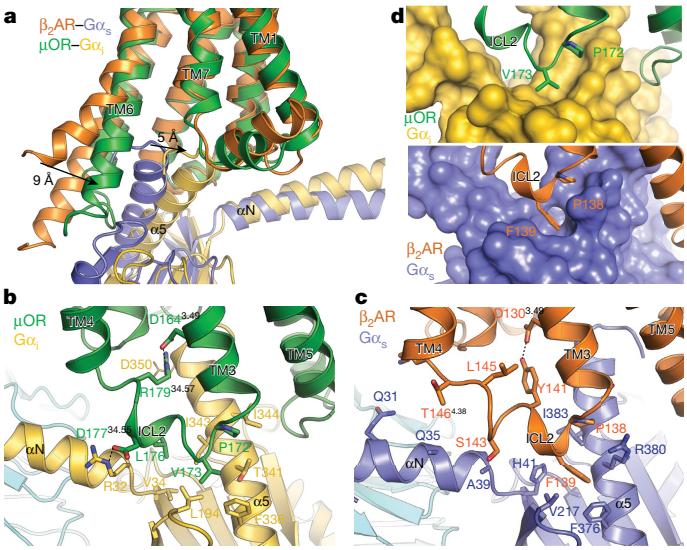


Fig. 4 | Comparison of the receptor–G-protein binding interfaces of the μOR–G_i and β_2AR–G_s complexes. a, Comparison of the conformation of the α5-helix of Gα and receptor TM6 in β_2AR–G_s and μOR–G_i complexes after alignment on the receptor. **b**, Interactions between ICL2 of the μOR (green) and $G\alpha_i$ (gold). Asp 350 of $G\alpha_i$ is depicted with narrow lines to indicate uncertainty in its conformation due to poor cryo-EM density for its side chain. **c**, Interactions between ICL2 of the β_2AR (orange) and $G\alpha_s$ (blue). **d**, Surface view of the hydrophobic pockets in $G\alpha_i$ (top panel) and $G\alpha_s$ (bottom panel) that interact with a non-polar amino acid in ICL2 of the μOR and β_2AR, respectively.

both complexes occur between ICL2, ICL3 and TM3, TM5 and TM6 on the receptor and the αN, αN–β1 loop and α5-helix on the Gα subunit of the G protein (Fig. 4). The most striking differences between the β_2AR–G_s and μOR–G_i complexes are in the relative position of the α5-helix of both G proteins and the corresponding shift in the position of TM6 in the receptor. The α5-helix of $G\alpha_i$ is rotated ~21° relative to the α5-helix of $G\alpha_s$, leading to a 5 Å displacement of the extreme C terminus of the $G\alpha_i$ helix α5 toward TM7 of the μOR (Fig. 4a). This difference in α5 positioning is associated with a smaller outward displacement of the μOR TM6. The C-terminal residues of α5 that interact with TM5 and TM6 of the receptor are bulkier in G_s than in G_i, with tyrosine and glutamate in place of cysteine and glycine at positions −4 and −3 from the C terminus, respectively. Accordingly, substitution of these two amino acids of G_s into G_i would lead to steric clashes with TM3 and the TM7–helix 8 loop (Extended Data Fig. 7). In the G_s-coupled family B calcitonin[11] and GLP-1[10] receptors, G-protein coupling is associated with a large kink of TM6 at the conserved PXXG motif, which produces an even larger outward displacement of TM6 than that observed in the β_2AR–G_s complex.

Surprisingly, the structure of μOR–G_i shows substantial similarity to an active-state structure of the visual pigment rhodopsin (metarhodopsin II) in complex with a modified peptide derived from the 11 C-terminal residues of the α-subunit of the visual G protein transducin ($G\alpha_t$CT2)[28] (Extended Data Fig. 8). Despite the absence of the rest of the heterotrimeric G protein in the metarhodopsin II–$G\alpha_t$CT2 structure, the conformation of TM6 of metarhodopsin II is highly similar to that of the μOR, and the location of the $G\alpha_t$CT2 peptide is almost identical to that of the C terminus of G_i in complex with μOR. This finding is consistent with observations showing that substitution of the last five amino acids of the α5-helix of Gα is sufficient to change G-protein-coupling specificity[29].

In Extended Data Table 2, we list amino acids in the μOR that interact with the cytoplasmic surface of G_i. The ICL2 of μOR primarily forms interactions with the αN and α5 helices of $G\alpha_i$, including a key ionic interaction between $D177^{34.55}$ of μOR (G Protein Coupled Receptor Data Base (GPCRDB) numbering[30]) in ICL2 and R32 in the αN–β1 loop of $G\alpha_i$ (Fig. 4b). Although $D^{34.55}$ in ICL2 is conserved in all opioid receptors with available sequences (GPCRDB[31]), it is variable in most other G_i-coupled receptors. Another notable interaction involves $R179^{34.57}$ in ICL2 of μOR, which simultaneously coordinates the highly conserved $D164^{3.49}$ in the $D^{3.49}R^{3.50}Y^{3.51}$ motif and potentially forms an additional interaction with D350 (−5 position) in the α5-helix of $G\alpha_i$ (Fig. 4b). This arginine residue is essential for μOR-induced G_i signalling, as the polymorphic variant R179C abolishes signalling in vitro[32] and leads to insensitivity to morphine in patients homozygous for the mutation[33]. The potential role of this interaction network in G-protein coupling is supported by the preponderance of basic residues (arginine and lysine) at this position in most G_i-coupled receptors, whereas G_s-coupled receptors contain alternative residues at the equivalent position (Extended Data Table 2).

A further group of contacts occurs between $P172^{34.50}$ and $V173^{34.51}$ of μOR and a hydrophobic patch on $G\alpha_i$ comprised of residues F336, I343, I344 and T340 on the α5-helix and L194 on the β2–β3 loop (Fig. 4b, d). In the GDP bound state, these α5-helix residues are buried by the adjacent β2 and β3 loops. Coupling to a receptor involves an upward shift of the α5-helix and exposes these residues to form a shallow hydrophobic pocket that interacts with μOR $V173^{34.51}$ in ICL2 (Fig. 4b, d). In the case of G_s, a deeper hydrophobic pocket in this region engages the bulky aromatic $F139^{34.51}$ in ICL2 of the β_2AR (Fig. 4c, d).

In the μOR, ICL3 stabilizes the interface between receptor and G protein through two sets of interactions: one set involves multiple contacts with a hydrophobic patch on the α5-helix of $G\alpha_i$, while another engages the β6 strand of $G\alpha_i$ through a network of charged residues (Fig. 5a, c). The hydrophobic interface formed by ICL3 is similar in both the μOR and β_2AR; in the β_2AR, TM5 is helically extended to form a larger hydrophobic interaction around nonpolar residues in the α5-helix of $G\alpha_s$ (Fig. 5b). The shorter ICL3 of the μOR does not form a similar helical extension, but it nevertheless fulfills the same role. Residues $V262^{5.68}$, M264 and L265 fold back to form a hydrophobic patch that interacts with hydrophobic residues on the α5-helix of $G\alpha_i$ (Fig. 5a).

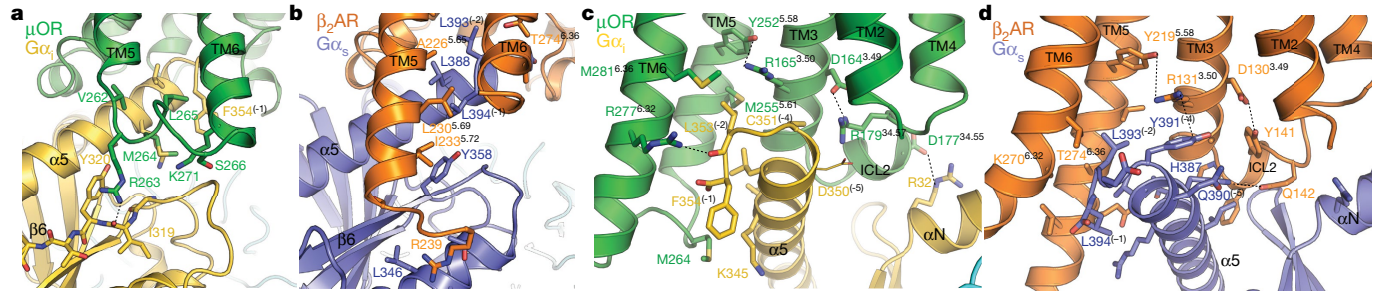


Fig. 5 | Comparison of the receptor–G-protein binding interfaces of the μOR–G_i and β_2AR–G_s complexes. a, **c**, Interactions between ICL3 of μOR and $G\alpha_i$ (**a**) and between the cytosolic ends of TM3, TM5 and TM6 of the μOR and the α5-helix of G_i (**c**). D350 of $G\alpha_i$ is depicted with narrow lines to indicate uncertainty in its position due to poor cryo-EM density for its side chain. **b**, **d**, The analogous interfaces between β_2AR and G_s (**b**, **d**).

The second set of polar contacts involves R263 of μOR and the backbone carbonyl of I319 on the β6 strand of $G\alpha_i$ (Fig. 5a). Mutations of R263 reduce, but do not abolish, G_i signalling[34], which is consistent with the potential importance of stabilizing the β6 strand of $G\alpha_i$ in the observed conformation. There is no analogous interaction in the β_2AR–G_s complex (Fig. 5b). This additional recognition interface may be necessary for efficient μOR–G_i coupling owing to the higher affinity for GDP of G_i relative to G_s. Compared to G_s-coupled receptors, additional interactions with the β6 strand in G_i-coupled receptors may be required to disrupt interactions between the Ras domain and GDP for efficient nucleotide exchange.

The cytosolic ends of μOR TM3, TM5 and TM6 further stabilize the nucleotide-free conformation of the α5-helix by interacting with highly conserved residues in the distal C terminus of $G\alpha_i$ (Fig. 5c). In particular, C351 (−4 position) of $G\alpha_i$ is in close proximity to the cytosolic end of TM3 of μOR. This cysteine residue has previously been identified as the site of pertussis toxin-mediated inhibition of $G_{i/o}$ family proteins by enzymatic ADP ribosylation[35]. The close apposition of C351 to the cytoplasmic surface of μOR highlights how the addition of a bulky modification at this position can completely inhibit receptor coupling and nucleotide exchange[35]. In addition to this interaction, μOR residues $M255^{5.61}$, $I278^{6.33}$, $M281^{6.36}$ and $V282^{6.37}$ form a hydrophobic pocket that engages the absolutely conserved $G\alpha_i$ residue L353 (−2 position) in the α5-helix. $M255^{5.61}$ and $M281^{6.36}$ have previously been observed in NMR experiments to respond to activation by DAMGO[36], suggesting that this region undergoes conformational changes prior to G-protein coupling. Further stabilization, however, is likely to be provided by a hydrogen bond between $R277^{6.32}$ and the backbone carbonyl of L353 (Fig. 5b). Notably, interactions between the C terminus of the α5-helix and the receptor core are entirely different in the β_2AR–G_s complex (Fig. 5d).

Our findings provide structural insights into why μOR does not couple to G_s, but do not explain the mechanism of G-protein-coupling specificity across all GPCRs. It is possible that coupling specificity is determined at an intermediate step in the formation of a GPCR–G-protein complex, such as the initial interactions between the GDP-bound G protein and the agonist-bound receptor. Recent single molecule fluorescence studies provide evidence for a transient intermediate complex between GDP-bound G_s and the β_2AR that is associated with a smaller outward movement of TM6[37]. Previous studies suggest that amino acids C-terminal to helix 8 confer coupling specificity for G_q in the M_3 muscarinic receptor (M_3R)[38]. Given that there are no interactions between the C termini of the β_2AR or μOR and their respective G proteins in the nucleotide-free complexes, we hypothesize that engagement of G_q with the C terminus of M_3R may occur at an earlier stage in complex formation. Thus, the nucleotide-free GPCR–G-protein complex may be preceded by one or more GDP-bound intermediates characterized by dynamic low-affinity interactions with the receptor. Such 'initial encounter' complexes may show larger energetic differences among interactions with various G-protein subtypes than the nucleotide-free state, and would thereby contribute more critically to coupling specificity. The transient nature of such interactions, however, poses challenges for structure determination by crystallography or cryo-EM.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0219-7.

Received: 31 December 2017; Accepted: 4 May 2018;
Published online 13 June 2018.

1. Matthes, H. W. et al. Loss of morphine-induced analgesia, reward effect and withdrawal symptoms in mice lacking the mu-opioid-receptor gene. *Nature* **383**, 819–823 (1996).
2. Barnett, M. L., Olenski, A. R. & Jena, A. B. Opioid-prescribing patterns of emergency physicians and risk of long-term use. *N. Engl. J. Med.* **376**, 663–673 (2017).
3. Connor, M. & Christie, M. D. Opioid receptor signalling mechanisms. *Clin. Exp. Pharmacol. Physiol.* **26**, 493–499 (1999).
4. Raffa, R. B., Martinez, R. P. & Connelly, C. D. G-protein antisense oligodeoxyribonucleotides and μ-opioid supraspinal antinociception. *Eur. J. Pharmacol.* **258**, R5–R7 (1994).
5. Raehal, K. M., Walker, J. K. L. & Bohn, L. M. Morphine side effects in β-arrestin 2 knockout mice. *J. Pharmacol. Exp. Ther.* **314**, 1195–1201 (2005).
6. Schmid, C. L. et al. Bias factor and therapeutic window correlate to predict safer opioid analgesics. *Cell* **171**, 1165–1175.e13 (2017).
7. DeWire, S. M. et al. A G protein-biased ligand at the μ-opioid receptor is potently analgesic with reduced gastrointestinal and respiratory dysfunction compared with morphine. *J. Pharmacol. Exp. Ther.* **344**, 708–717 (2013).
8. Manglik, A. et al. Structure-based discovery of opioid analgesics with reduced side effects. *Nature* **537**, 185–190 (2016).
9. Rasmussen, S. G. F. et al. Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549–555 (2011).
10. Zhang, Y. et al. Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature* **546**, 248–253 (2017).
11. Liang, Y.-L. et al. Phase-plate cryo-EM structure of a class B GPCR–G-protein complex. *Nature* **546**, 118–123 (2017).
12. Huang, W. *et al.* Structural insights into μ-opioid receptor activation. *Nature* **524**, 315–321 (2015).
13. Ballesteros, J. A. & Weinstein, H. in *Receptor Molecular Biology* Vol. 25 (ed. Sealfon, S. C.) Ch. 19 (Elsevier, 1995).
14. Manglik, A. et al. Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature* **485**, 321–326 (2012).
15. Granier, S. et al. Structure of the δ-opioid receptor bound to naltrindole. *Nature* **485**, 400–404 (2012).
16. Wu, H. et al. Structure of the human κ-opioid receptor in complex with JDTic. *Nature* **485**, 327–332 (2012).
17. Che, T. et al. Structure of the nanobody-stabilized active state of the kappa opioid receptor. *Cell* **172**, 55–67.e15 (2018).
18. Fenalti, G. et al. Structural basis for bifunctional peptide recognition at human δ-opioid receptor. *Nat. Struct. Mol. Biol.* **22**, 265–268 (2015).
19. Emmerson, P. J., Liu, M. R., Woods, J. H. & Medzihradsky, F. Binding affinity and selectivity of opioids at mu, delta and kappa receptors in monkey brain membranes. *J. Pharmacol. Exp. Ther.* **271**, 1630–1637 (1994).
20. Minami, M. et al. DAMGO, a μ-opioid receptor selective ligand, distinguishes between μ- and κ-opioid receptors at a different region from that for the distinction between μ- and δ-opioid receptors. *FEBS Lett.* **364**, 23–27 (1995).
21. DiMaio, J. & Schiller, P. W. A cyclic enkephalin analog with high in vitro opiate activity. *Proc. Natl Acad. Sci. USA* **77**, 7162–7166 (1980).
22. Rasmussen, S. G. F. et al. Structure of a nanobody-stabilized active state of the β_2 adrenoceptor. *Nature* **469**, 175–180 (2011).
23. Wall, M. A. et al. The structure of the G protein heterotrimer $G_{i\alpha1}\beta_1\gamma_2$. *Cell* **83**, 1047–1058 (1995).
24. Dror, R. O. et al. Signal transduction. Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science* **348**, 1361–1365 (2015).
25. Van Eps, N. et al. Interaction of a G protein with an activated receptor opens the interdomain interface in the alpha subunit. *Proc. Natl Acad. Sci. USA* **108**, 9420–9424 (2011).
26. Kaya, A. I. et al. A conserved phenylalanine as a relay between the α5 helix and the GDP binding region of heterotrimeric G_i protein α subunit. *J. Biol. Chem.* **289**, 24475–24487 (2014).
27. Chung, K. Y. et al. Conformational changes in the G protein Gs induced by the β_2 adrenergic receptor. *Nature* **477**, 611–615 (2011).
28. Choe, H. W. *et al.* Crystal structure of metarhodopsin II. *Nature* **471**, 651–655 (2011).
29. Conklin, B. R., Farfel, Z., Lustig, K. D., Julius, D. & Bourne, H. R. Substitution of three amino acids switches receptor specificity of $G_q\alpha$ to that of $G_i\alpha$. *Nature* **363**, 274–276 (1993).
30. Isberg, V. et al. Generic GPCR residue numbers—aligning topology maps while minding the gaps. *Trends Pharmacol. Sci.* **36**, 22–31 (2015).
31. Isberg, V. et al. GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **42**, D422–D425 (2014).
32. Ravindranathan, A. et al. Functional characterization of human variants of the mu-opioid receptor gene. *Proc. Natl Acad. Sci. USA* **106**, 10811–10816 (2009).
33. Skorpen, F. et al. The rare Arg181Cys mutation in the μ opioid receptor can abolish opioid responses. *Acta Anaesthesiol. Scand.* **60**, 1084–1091 (2016).
34. Chaipatikul, V., Loh, H. H. & Law, P. Y. Ligand-selective activation of μ-opioid receptor: demonstrated with deletion and single amino acid mutations of third intracellular loop domain. *J. Pharmacol. Exp. Ther.* **305**, 909–918 (2003).
35. West, R. E., Moss, J., Vaughan, M., Liu, T. & Liu, T. Y. Pertussis toxin-catalyzed ADP-ribosylation of transducin. Cysteine 347 is the ADP-ribose acceptor site. *J. Biol. Chem.* **260**, 14428–14430 (1985).
36. Okude, J. et al. Identification of a conformational equilibrium that determines the efficacy and functional selectivity of the μ-opioid receptor. *Angew. Chem. Int. Edn Engl.* **54**, 15771–15776 (2015).
37. Gregorio, G. G. et al. Single-molecule analysis of ligand efficacy in β_2AR-G-protein activation. *Nature* **547**, 68–73 (2017).
38. Qin, K., Dong, C., Wu, G. & Lambert, N. A. Inactive-state preassembly of G_q-coupled receptors and G_q heterotrimers. *Nat. Chem. Biol.* **7**, 740–747 (2011).

Acknowledgements We thank J.-P. Carralot (F. Hoffmann–La Roche) for help in antibody generation, C. Yoshioka and C. Lopez for assistance with data collection, M. Siegrist, G. Schmid, B. Rutten, D. Zulauf, S. Kueng (Roche Non-Clinical Biorepository) and R. Thoma for technical assistance with biomass

and cell line generation. We also acknowledge N. Moriarty (Lawrence Berkeley National Laboratories) for help with generation of parameters for DAMGO and general advice for refinement of our model. The work is supported by NIH grant R37DA036246 (B.K.K., S.G. and G.S.) and NIH grant R01GM083118 (B.K.K.). Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number T32GM007276 (A.K.). The content is solely the responsibility of the authors and does not necessarily represent the official view of the National Institutes of Health. S.M. was supported by the Roche Postdoctoral Fellowship (RPF ID: 113). G.F.X.S. acknowledges the Swiss National Science Foundation for grants 310030_153145 and 310030B_17335 and long-term financial support from the Paul Scherrer Institute. B.K.K. is a Chan–Zuckerberg Biohub Investigator.

Reviewer information *Nature* thanks L. Shi, D. Wacker and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.K. prepared the μOR–G_i complex and refined the structure from cryo-EM density maps. H.H. obtained and processed cryo-EM data with the assistance of Y.Z. and Q.Q. S.M. identified and prepared scFv16 with assistance from R.D. and H.M and under supervision of G.F.X.S. A.M., D.H. S.G. and A.K. developed the procedure for forming the μOR–G_i complex. N.R.L and J.M.P performed molecular dynamics simulations under supervision of R.O.D. W.I.W. aided in map interpretation and model refinement. A.K., A.M., B.K.K. and G.S. wrote the manuscript. A.M., G.S. and B.K.K. supervised the project.

Competing interests B.K. is a founder of and consultant for ConfometRx.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0219-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0219-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.M. or G.S. or B.K.K.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Purification of μOR. These studies utilized a previously described mouse μOR construct with cleavable N- and C-terminal domains[12]. In brief, the receptor was expressed in *Spodoptera frugiperda* Sf9 insect cells using the baculovirus method (Expression Systems), extracted from insect cell membranes with *n*-dodecyl-β-D-maltoside (DDM, Anatrace), and purified by nickel-chelating sepharose chromatography. The Ni-NTA eluate was loaded onto M1 anti-Flag immunoaffinity resin and washed with progressively lower concentrations of the antagonist naloxone. The μOR was then eluted in a buffer consisting of 20 mM Hepes pH 7.5, 100 mM NaCl, 0.1% DDM and 0.01% cholesterol hemisuccinate (CHS) supplemented with 50 nM naloxone, Flag peptide and 5 mM EDTA. The monomeric fraction was purified by size exclusion chromatography on a Superdex 200 10/300 gel filtration column (GE Healthcare) in 20 mM Hepes pH 7.5, 100 mM NaCl, 0.1% DDM, 0.01% CHS, and 1 μM DAMGO. A further twofold molar excess of DAMGO was added to the preparation and the resulting agonist-bound μOR preparation was concentrated to ~100 μM.

Expression and purification of heterotrimeric G_i. Heterotrimeric G_i was expressed and purified as previously described[24]. In brief, *Trichuplusia ni* Hi5 insect cells were co-infected with two viruses, one encoding the wild-type human $G\alpha_i$ subunit and another encoding the wild-type human $\beta_1\gamma_2$ subunits with an octahistidine tag inserted at the amino terminus of the β_1 subunit. Cultures were harvested 48 h post infection. Cells were lysed in hypotonic buffer and lipid-modified heterotrimeric G_i was extracted in a buffer containing 1% sodium cholate. The soluble fraction was purified using Ni-NTA chromatography, and the detergent was exchanged from cholate to DDM on a column. After elution, the protein was dialyzed against a buffer containing 20 mM Hepes pH 7.5, 100 mM NaCl, 0.015% DDM, 100 μM TCEP, 10 μM GDP, and concentrated to ~20 mg/ml for further complexing with the μOR.

Generation of scFv16. Six to eight week old female Balb/c mice were immunized with a purified rhodopsin–G_i complex[39]. Hybridoma cells were prepared using splenocytes of immunized mice using standard methods in combination with PAI myeloma cells (RRID: CVCL J288). Clones that showed a positive reaction to purified rhodopsin(N2C/D282C/M257Y)–G_{i1} complex in an ELISA assay and by immunoprecipitation were further characterized as monoclonal antibodies or Fab fragments. Fab-16 was selected from the initial pool of clones because it prevented dissociation of the rhodopsin(N2C/D282C/M257Y)–G_{i1} complex by GTPγS, and therefore acted as a stabilizing chaperone in the same manner as Nb35 for G_s. The full sequences of constructs used are listed in Supplementary Fig. 1. All animal studies were performed at Roche Innovation Center in Basel according to ethical guidelines and internal IRB approval. All cell lines were obtained from the manufacturer and tested for contamination.

A C-terminal hexahistidine-tagged single chain construct of Fab16 (scFv16) was cloned into a modified pVL1392 vector containing a GP67 secretion signal immediately prior to the amino terminus of the scFv, expressed in secreted form from *Trichuplusia ni* Hi5 insect cells using the baculovirus method, and purified by Ni-NTA chromatography. Supernatant from baculovirus-infected cells was pH balanced by addition of Tris pH 8.0. Chelating agents were quenched by addition of 1 mM nickel chloride and 5 mM calcium chloride and incubation with stirring for 1 h at 25 °C. Resulting precipitates were removed by centrifugation and the supernatant was loaded onto Ni-NTA resin. The column was washed with 20 mM Hepes pH 7.5, 500 mM NaCl, and 10 mM imidazole followed by a low salt wash comprised of the same buffer substituted with 100 mM NaCl. Following elution with the same buffer supplemented with 250 mM imidazole, the C-terminal hexahistidine tag was cleaved by incubation with human rhinovirus 3C protease, and the protein was dialyzed into a buffer consisting of 20 mM Hepes pH 7.5 and 100 mM NaCl. Cleaved scFv16 was further purified by reloading over Ni-NTA resin. The flow-through was collected and purified over gel filtration chromatography using a Superdex 200 16/60 column. Monomeric fractions were pooled, concentrated, and flash frozen in liquid nitrogen until further use.

Formation and purification of the μOR–G_i–scFv16 complex. Purified DAMGO-bound μOR was mixed with a 1.2 molar excess of G_i heterotrimer. The coupling reaction was allowed to proceed at 24 °C for 1 h and was followed by addition of apyrase to catalyze hydrolysis of unbound GDP, which destabilizes the nucleotide-free complex[40]. After one more hour at 25 °C, a fourfold volume of 20 mM Hepes pH 7.5, 100 mM NaCl, 1% lauryl maltose neopentyl glycol (L-MNG), 0.1% CHS was added to the complexing reaction to initiate detergent exchange. After 1 h incubation at 25 °C to allow micelle exchange, 1 mM $MnCl_2$ and lambda phosphatase (New England Biolabs) were added to dephosphorylate the preparation. This reaction was further incubated at 4 °C for 2 h. To remove excess G protein and residual DDM, the complexing mixture was purified by M1 anti-Flag affinity chromatography. Bound complex was first washed in a buffer containing 1% L-MNG, followed by washes in gradually decreasing L-MNG concentrations. The complex was then eluted in 20 mM Hepes pH 7.5, 100 mM NaCl, 0.01% L-MNG, 0.001% CHS, 300 nM DAMGO, 5 mM EDTA and Flag peptide. The eluted complex was supplemented with 100 μM TCEP to provide a reducing environment. The tobacco etch virus (TEV) protease and human rhinovirus 3C protease were added to cleave the flexible μOR N and C termini. Finally, a 1.2 molar excess of scFv16 was added to the preparation. Once cleavage of the termini was confirmed by SDS–PAGE, the μOR–G_i–scFv16 complex was purified by size exclusion chromatography on a Superdex 200 10/300 column in 20 mM Hepes pH 7.5, 100 mM NaCl, 300 nM DAMGO, 0.00075% L-MNG and 000025% glyco-diosgenin (GDN) with 0.0001% CHS total. Peak fractions were concentrated to ~7 mg/ml for electron microscopy studies.

Cryo-EM and 3D reconstructions of μOR–G_i–scFv16 complex. Three microlitres of purified μOR–G_i–scFv16 complex was applied to glow-discharged 200-mesh grids (Quantifoil R1.2/1.3) and subsequently vitrified using a Vitrobot Mark IV (Thermo Fischer Scientific). Cryo-EM imaging was performed on a Titan Krios operated at 300 kV at a nominal magnification of 130,000× using a Gatan K2 Summit direct electron camera in counted mode, corresponding to a pixel size of 1.04 Å. A total of 2642 image stacks were obtained with a defocus range of −0.8 to −2.6 μm. Each stack movie was recorded for a total of 8 s with 0.1 s per frame. The dose rate was 5 $e^-/Å^2/s$, resulting in an accumulated dose of 40 electrons per $Å^2$.

Dose-fractionated image stacks were subjected to beam-induced motion correction using MotionCor2[41]. A sum of all frames, filtered according to exposure dose, in each image stack was used for further processing. CTF parameters for each micrograph were determined by Gctf v1.06[42]. Particle selection, two-dimensional and three-dimensional classification, and 3D reconstruction were performed using RELION2.1[43], apart from the last round of local refinement and reconstruction that was performed with Frealign[44]. Semi-automated selected 893,426 particle projections were subjected to reference-free 2D classification and averaging using a binned dataset with a pixel size of 2.08 Å. Particles (379,373) with well-defined averages were subjected to further processing. An ab initio map generated by VIPER[45] was used as initial reference model for maximum-likelihood-based 3D classification, which, however, did not produce classes with notable differences. Thus, all 379,373 particle projections were subjected to 3D refinement, producing a map at 4.3 Å resolution. The dataset was further reduced by removing particle projections from micrographs with resolution lower than 4.5 Å, resulting in a dataset of 359,406 particles that were subjected to refinement and reconstruction after subtracting densities for the mobile Gα α-helical domain and the detergent micelle[11]. Particle projection assignments from RELION were imported into Frealign[46] for a final round of local refinement and reconstruction. To prevent overfitting, the resolution limit for each alignment iteration never exceeded the 0.9 value of the Frealign calculated Fourier shell correlation (FSC). The map was further improved after additionally subtracting densities corresponding to the ScFv from the raw particle projections[11]. The indicated resolution, using Phenix 'gold standard' FSC[47], of the final reconstruction is 3.5 Å and 3.6 Å at FSC 0.143 for the ScFv-subtracted map and the ScFv-including map, respectively. Local resolution was determined using the Bsoft package[48] with unfiltered half maps as input.

Model building and refinement. The building of a full atomic model for the μOR–G_i complex was aided by the quality and resolution of our map, as well as the existence of high-resolution crystal structures of each of the components that make up the complex. A composite model was formed by rigid body fitting of the active-state μOR (PDB code 5C1M)[12] with nanobody removed, as well as the Ras domain and βγ subunits of GDP-bound G_i (PDB 1GP2)[23]. The α5-helix of $G\alpha_i$ was removed and manually fitted to the density, and the final eight residues that were missing from the extreme C terminus of the 1GP2 structure were manually built in Coot[49]. This starting model was then subjected to iterative rounds of automated refinement in Rosetta[50] and Phenix real space refine[47], and manual building in Coot[49]. In the regions of the model for which side-chain density was too weak to unambiguously assign a conformation, we stubbed residues to their C_β position, while preserving sequence information (Supplementary Fig. 2, 3). The final model was visually inspected for general fit to the map, and geometry was further evaluated using Molprobity[51] as part of the Phenix suite of software. Initial restraints for DAMGO were generated using the PRODRG server[52]. To further refine the pose of DAMGO, we chose a pose from molecular dynamics simulation consistent with our map and then performed a refinement using Phenix. This involved manually editing the residue and atom names from a CHARMM parameter file to match the three-letter codes and atom names from the RCSB. In particular, DAL for D-alanine, MEA for *N*-methyl phenylalanine, and ETA for Gly-ol C terminus. Additional, custom, restraints were generated to maintain planarity of the final peptide bond between MEA and ETA as a supplement to the natural library of Phenix amino acid restraints. Model overfitting was evaluated through its refinement against one cryo-EM half map after randomly displacing all atoms by 0.2 Å. FSC curves were calculated between the resulting model and the half map used for refinement (red curve, Extended Data Fig. 3b, c), as well as between the resulting

model and the other half map for cross validation (green curve, Extended Data Fig. 3b, c), and also against the full map (black curve, Extended Data Fig. 3b, c). The final refinement statistics for both models are provided in Extended Data Table 1.

System setup for molecular dynamics simulations. Molecular dynamics simulations were initiated from an earlier refinement of the structure reported in this study after removing the G protein and ScFv fragment. Prior to beginning simulations, Schrödinger Glide[53] was used to relax DAMGO to an energetically favourable conformation. The initial DAMGO pose is depicted in Extended Data Fig. 3. We performed five independent simulations, for each of which initial atom velocities were assigned randomly and independently. Prime (Schrödinger) was used to model missing side chains, and neutral acetyl and methylamide groups were added to cap protein termini. Titratable residues remained in their dominant protonation state at pH 7, as determined using PropKa, except for $D^{2.50}$ and $D^{3.49}$, which were protonated. Our simulations incorporated the waters from the 5C1M crystal structure.

The prepared protein structures were aligned to the 'orientation of proteins in membranes' (OPM) structure for PDB entry 5C1M[54]. The aligned structures were then inserted into a pre-equilibrated palmitoyl-oleoyl-phosphatidylcholine (POPC) bilayer using Dabble, a simulation preparation software[55]. Sodium and chloride ions were added to neutralize each system at a concentration of approximately 150 mM. Bilayer dimensions were chosen to maintain at least a 30 Å buffer between protein images in the *x*–*y* plane and a 20 Å buffer between protein images in the *z* direction. Final system dimensions were approximately $80 \times 75 \times 90$ Å^3. Simulation times for each replicate were approximately 1 μs.

Molecular dynamics simulation protocols. We used the CHARMM36m force field for proteins, lipids and ions and the TIP3P model for waters[56–60]. Parameters for the non-canonical residues in DAMGO were determined by analogy to *N*-methyl glycine for assigning *N*-methyl parameters to *N*-methyl phenylalanine (residue 4) and by analogy to serine to assign parameters to the Gly-ol capping group (residue 5). CMAP terms for D-alanine were inverted from those for L-Alanine to account for the inverted chirality of the residue.

We performed the simulations using the Compute Unified Device Architecture (CUDA) version of Particle-Mesh Ewald Molecular Dynamics (PMEMD) in AMBER on one or two graphical processing units (GPUs)[61]. Simulations were performed using the AMBER16[62] software. Three rounds of minimization were performed, each consisting of 500 iterations of steepest descent minimization, followed by 500 iterations of conjugate gradient descent minimization, with harmonic restraints of 10.0, 5.0 and 1.0 kcal mol^{-1} Å^{-2} placed on the protein and lipids. Systems were heated from 0 K to 100 K in the NVT ensemble over 12.5 ps and then from 100 K to 310 K in the NPT ensemble over 125 ps, using 10.0 kcal mol^{-1} Å^{-2} harmonic restraints applied to lipid and protein heavy atoms. Systems were then equilibrated at 310 K in the NPT ensemble at 1 bar, with harmonic restraints on all protein heavy atoms tapered off by 1.0 kcal mol^{-1} Å^{-2} starting at 5.0 kcal mol^{-1} Å^{-2} in a stepwise fashion every 2 ns for 10 ns and then by 0.1 kcal mol^{-1} Å^{-2} in a stepwise fashion every 2 ns for 20 ns. Production simulations were performed in the NPT ensemble at 310 K and 1 bar, using a Langevin thermostat for temperature coupling and a Monte Carlo barostat for pressure coupling. These simulations used a 4-fs time step with hydrogen mass repartitioning[63]. Bond lengths to hydrogen atoms were constrained using SHAKE. Simulations used periodic boundary conditions. Non-bonded interactions were cut off at 9.0 Å, and long-range electrostatic interactions were computed using Particle Mesh Ewald (PME) with an Ewald coefficient of approximately 0.31 Å and an interpolation order of four. The FFT grid size was chosen such that the width of a grid cell was approximately 1 Å.

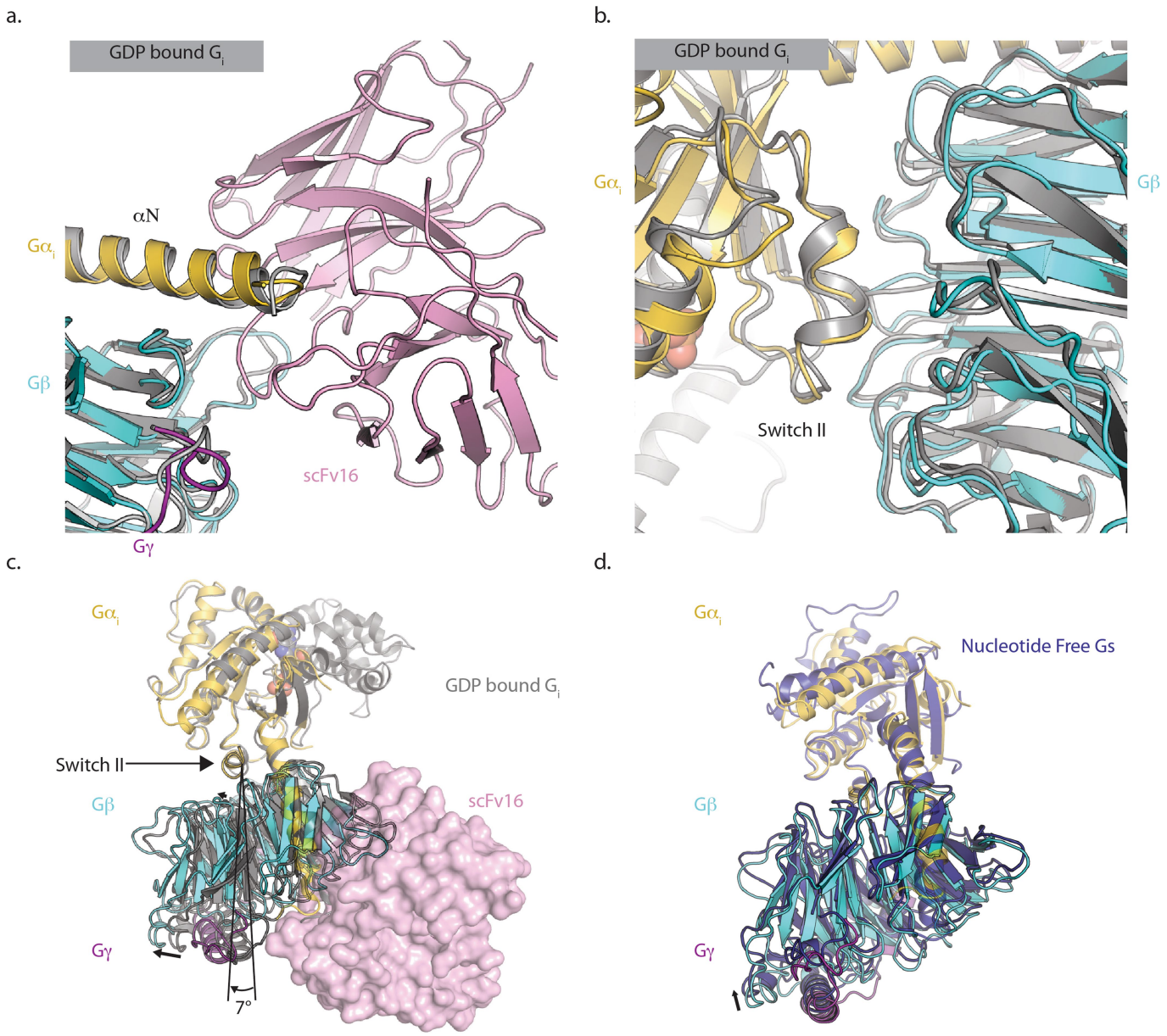
During production simulations, all residues within 5 Å of the G-protein interface were restrained to the initial structure using 5.0 kcal mol^{-1} Å^{-2} harmonic restraints applied to non-hydrogen atoms. Using such restraints reduces the overall system size, enabling more simulation, while ensuring that the receptor maintains an active conformation throughout the simulation.

Analysis protocols for molecular dynamics simulation. Trajectory snapshots were saved every 200 ps during production simulations. The AmberTools17 CPPTRAJ package was used to reimage and center trajectories[64]. Simulations were visualized and analysed using Visual Molecular Dynamics (VMD)[65]. In two simulations, DAMGO was trapped in an unstable binding pose, wherein the water-mediated interaction between the DAMGO tyrosine residue and H297 failed to form during equilibration, and instead a direct hydrogen bond between these residues was formed. Our analysis is based on the other three simulations, in which DAMGO's pose was consistent with the cryo-EM density. Water occupancy maps were generated using AmberTools17 GIST[66,67]. Frames from every 1 ns of simulation, excluding the first 400 ns, aligned to the initial structure, were used as input. The grid size was set to 0.25 Å. The resulting map was smoothed using a Gaussian filter with a standard deviation of two grid cells.

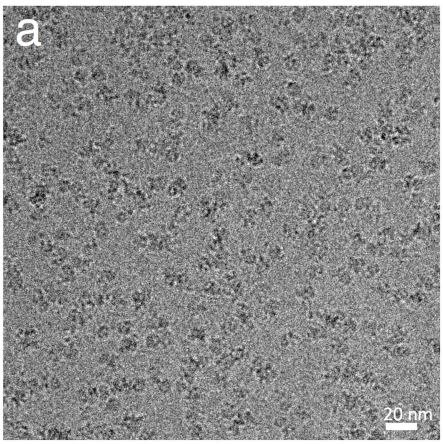
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

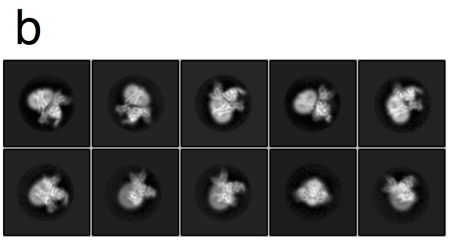
Data availability. All data generated or analysed during this study are included in this published article and its Supplementary Information. Sequences of constructs used in this study are listed in Supplementary Fig. 1. The cryo-EM density maps for the μOR–G_i complex with, and without, scFv16 have been deposited in the Electron Microscopy Data Bank (EMDB) under accession codes EMD-7868 and EMD-7869, respectively. The coordinates for the models of μOR–G_i with, and without, scFv-16 have been deposited in the Protein Data Bank (PDB) under accession numbers 6DDE and 6DDF respectively.

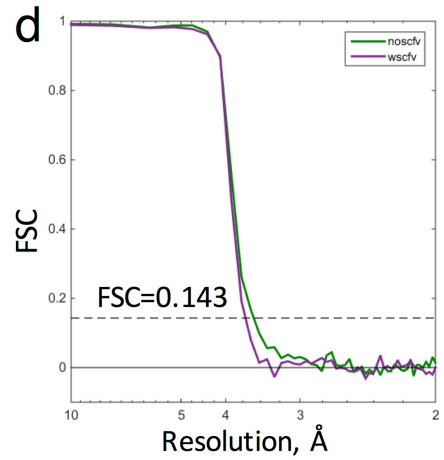
39. Maeda, S. et al. Crystallization scale preparation of a stable GPCR signaling complex between constitutively active rhodopsin and G-protein. *PLoS One* **9**, e98714 (2014).
40. Westfield, G. H. et al. Structural flexibility of the Gαs α-helical domain in the β2-adrenoceptor Gs complex. *Proc. Natl Acad. Sci. USA* **108**, 16086–16091 (2011).
41. Zheng, S. Q., Palovcak, E., Armache, J.-P., Verba, K. A., Cheng, Y. & Agard, D. A. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2016).
42. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
43. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination *J. Struct. Biol.* **180**, 519–530 (2012).
44. Grigorieff, N. FREALIGN: high-resolution refinement of single particle structures. *J. Struct. Biol.* **157**, 117–125 (2007).
45. Penczek, P. A., Grassucci, R. A. & Frank, J. The ribosome at improved resolution: new techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles. *Ultramicroscopy* **53**, 251–270 (1994).
46. Grigorieff, N. Frealign: an exploratory tool for single-particle Cryo-EM. *Methods Enzymol.* **579**, 191–226 (2016).
47. Adams, P. D. et al. The Phenix software for automated determination of macromolecular structures. *Methods* **55**, 94–106 (2011).
48. Heymann, J. B. & Belnap, D. M. Bsoft: image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157**, 3–18 (2007).
49. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
50. Wang, R. Y.-R. et al. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *eLife* **5**, 352 (2016).
51. Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
52. Schüttelkopf, A. W. & van Aalten, D. M. F. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D* **60**, 1355–1363 (2004).
53. Friesner, R. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
54. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinformatics* **22**, 623–625 (2006).
55. Betz, R. Dabble (v.2.6.3). https://doi.org/10.5281/zenodo.836914 (2017).
56. Best, R. B., Mittal, J., Feig, M. & MacKerell, A. D. Inclusion of many-body effects in the additive CHARMM protein CMAP potential results in enhanced cooperativity of α-helix and β-hairpin formation. *Biophys. J.* **103**, 1045–1051 (2012).
57. Best, R. B. et al. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ1 and χ2 dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).
58. Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145 (2013).
59. Klauda, J. B. et al. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* **114**, 7830–7843 (2010).
60. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2016).
61. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888 (2013).
62. Case, D. A. et al. Amber (v.16). http://ambermd.org (2018).
63. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-time-step molecular dynamics through hydrogen mass repartitioning. *J. Chem. Theory Comput.* **11**, 1864–1874 (2015).
64. Roe, D. R. & Cheatham, T. E. III. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
65. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
66. Nguyen, C. N., Young, T. K. & Gilson, M. K. Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit[7] uril. *J. Chem. Phys.* **137**, 044101 (2012).
67. Nguyen, C., Gilson, M. K. & Young, T. Structure and thermodynamics of molecular hydration via grid inhomogeneous solvation theory. Preprint at https://arXiv.org/abs/1108.4876 (2011).

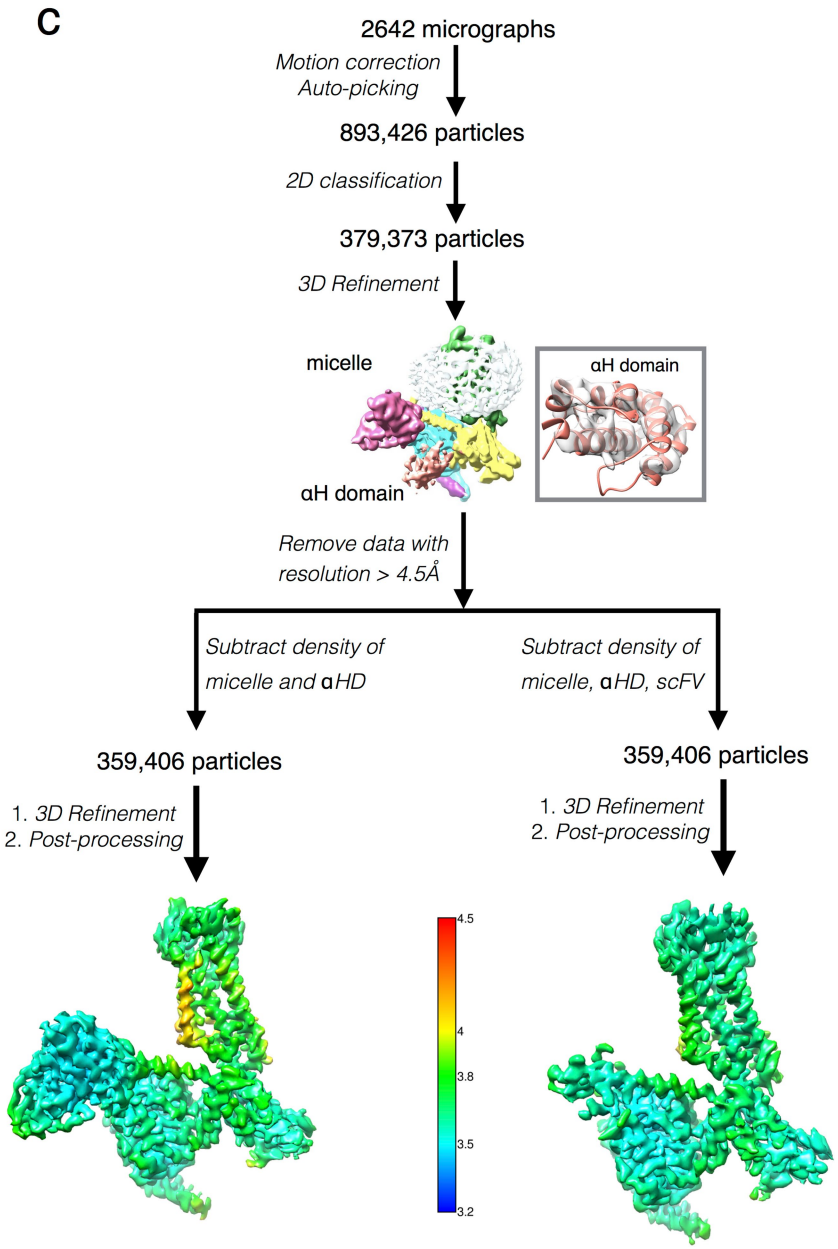


Extended Data Fig. 1 | Binding characteristics of scFv. a, **b**, scFv 16 does not perturb the interfaces between Gα and Gβ at its binding epitope (**a**) or the switch II region located ~40 Å away (**b**). Our structure is coloured by chain, whereas the structure of GDP-bound G_{i1} heterotrimer (PDB code 1GP2) is coloured grey. **c**, In the nucleotide-free state (coloured by subunit), there is a ~7° rotation of Gβγ relative to the $G\alpha_s$ switch II domain when compared to the GDP-bound form. **d**, This rotated conformation is similar to that observed in nucleotide-free G_s coupled to the β_2AR (PDB code 3SN6).

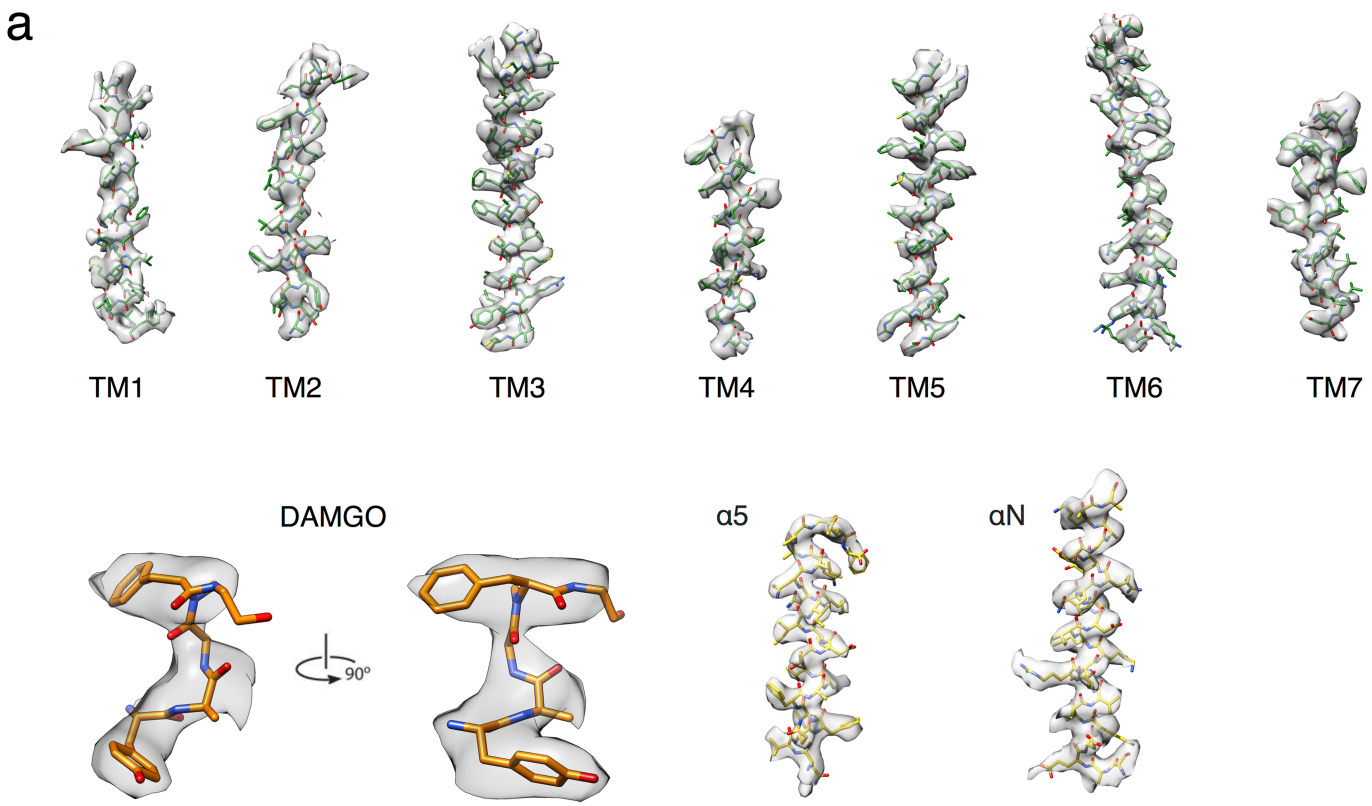


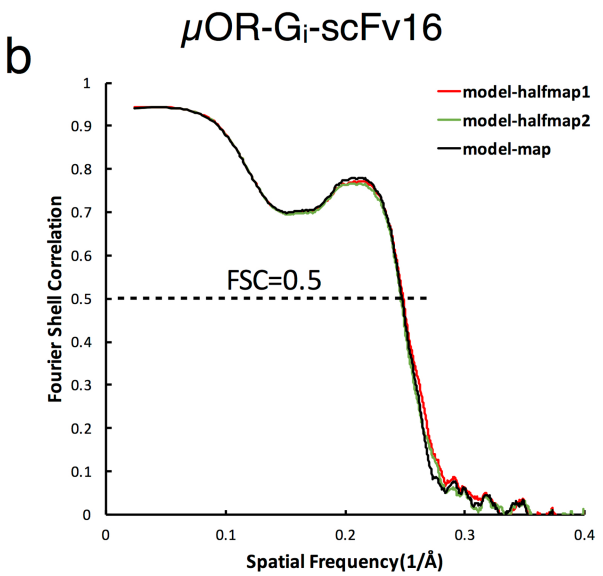


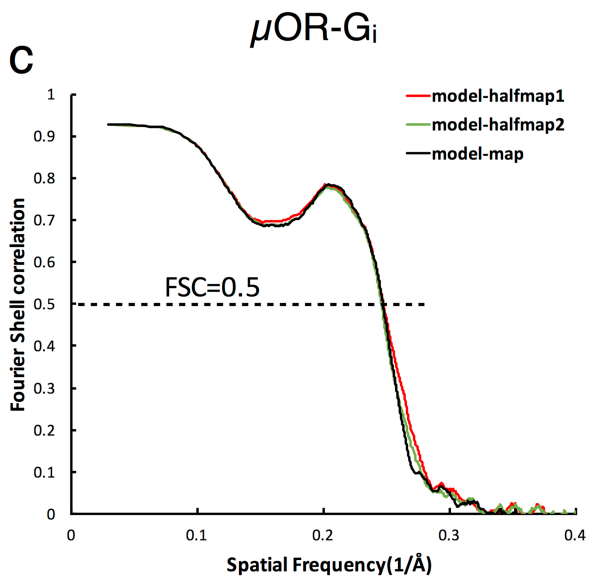




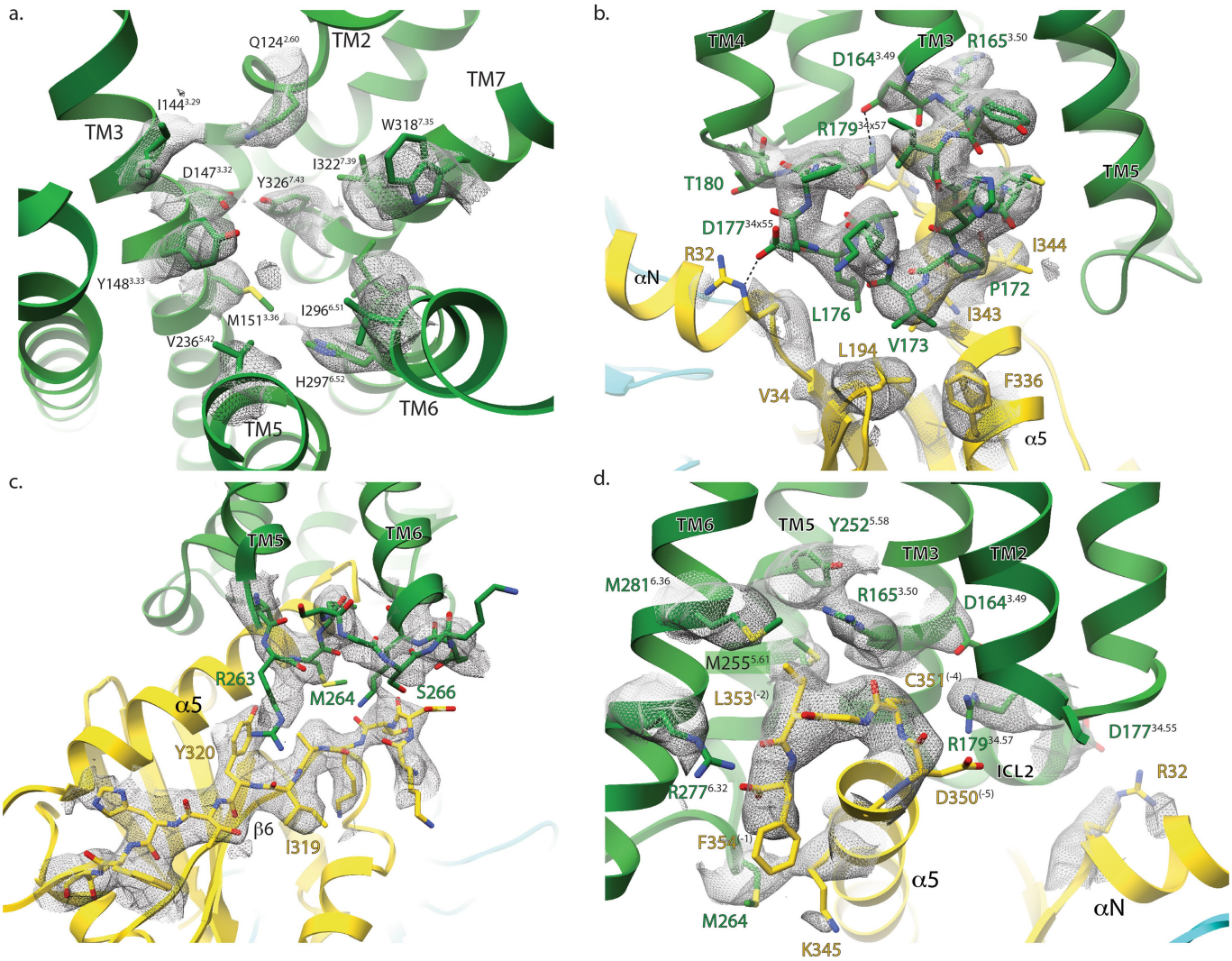
Extended Data Fig. 2 | Cryo-EM data processing. a, Representative cryo-EM image of the μOR–G_i complex. Scale bar, 20 nm. **b**, Representative 2D averages showing distinct secondary structure features from different views of the complex. **c**, Flow chart of cryo-EM data processing. The unmasked map in the middle of the chart has been coloured by subunit. The inset shows the fit of the crystal structure of the α-helical domain in the corresponding density of the unmasked reconstruction. 3D density maps coloured according to local resolution. **d**, 'Gold standard' FSC curves from Phenix indicate overall nominal resolutions of 3.5 Å and 3.6 Å using the FSC = 0.143 criterion for the scFv-subtracted map (green curve) and scFv-retained maps (purple curve), respectively.



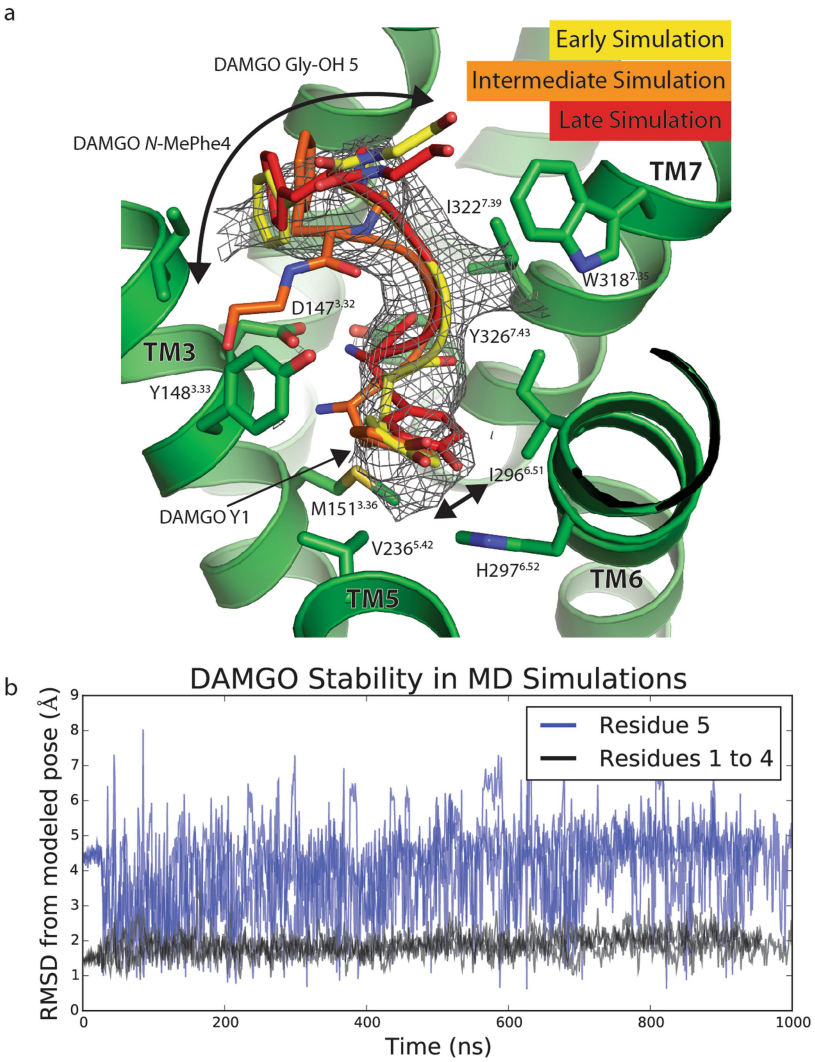




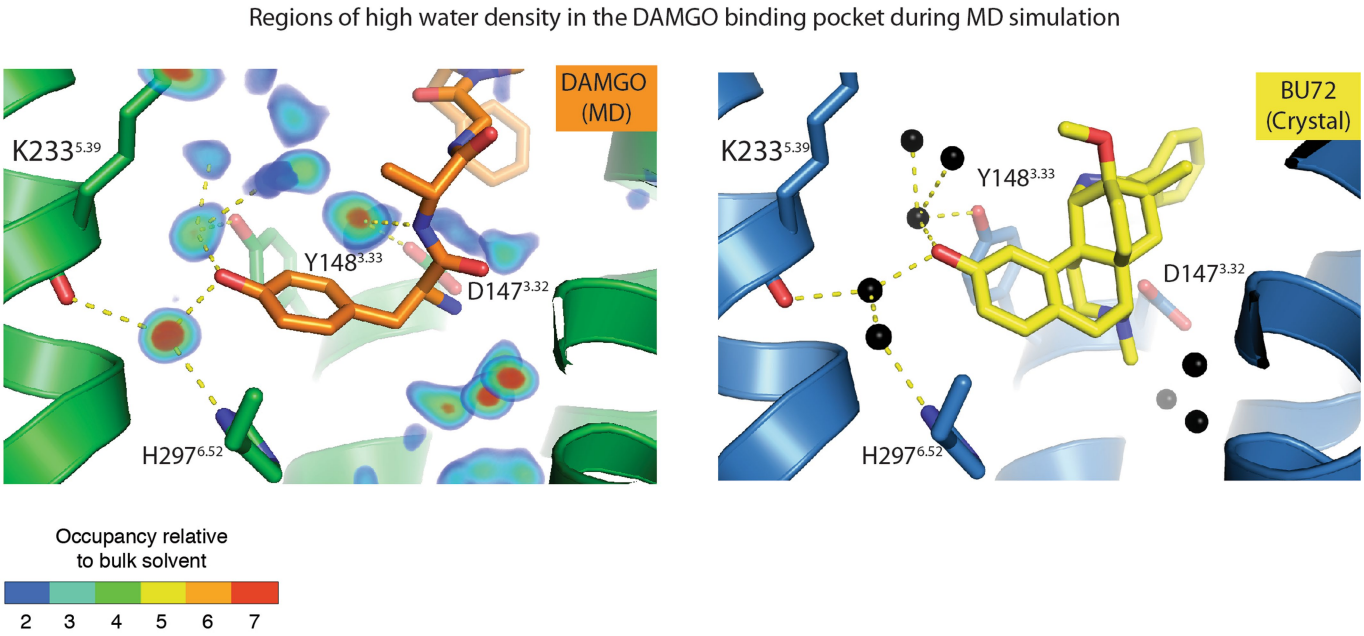
Extended Data Fig. 3 | Cryo-EM map versus refined structure. **a**, Cryo-EM density map (scFv subtracted) and model are shown for all seven transmembrane α-helices of the μOR, DAMGO, and Gα helices α5 and αN. **b**, **c**, Cross-validation of model to cryo-EM density map. The model was refined against one half map after displacement of atoms by 0.2 Å, and FSC curves were calculated between this model and the final cryo-EM map (full dataset, black) of the outcome of model refinement with a half map versus the same map (red), and of the outcome of model refinement with a half map versus the other half map (green). The results of the scFv-retained model versus map (**b**) and of scFv subtracted model versus map (**c**) are shown.



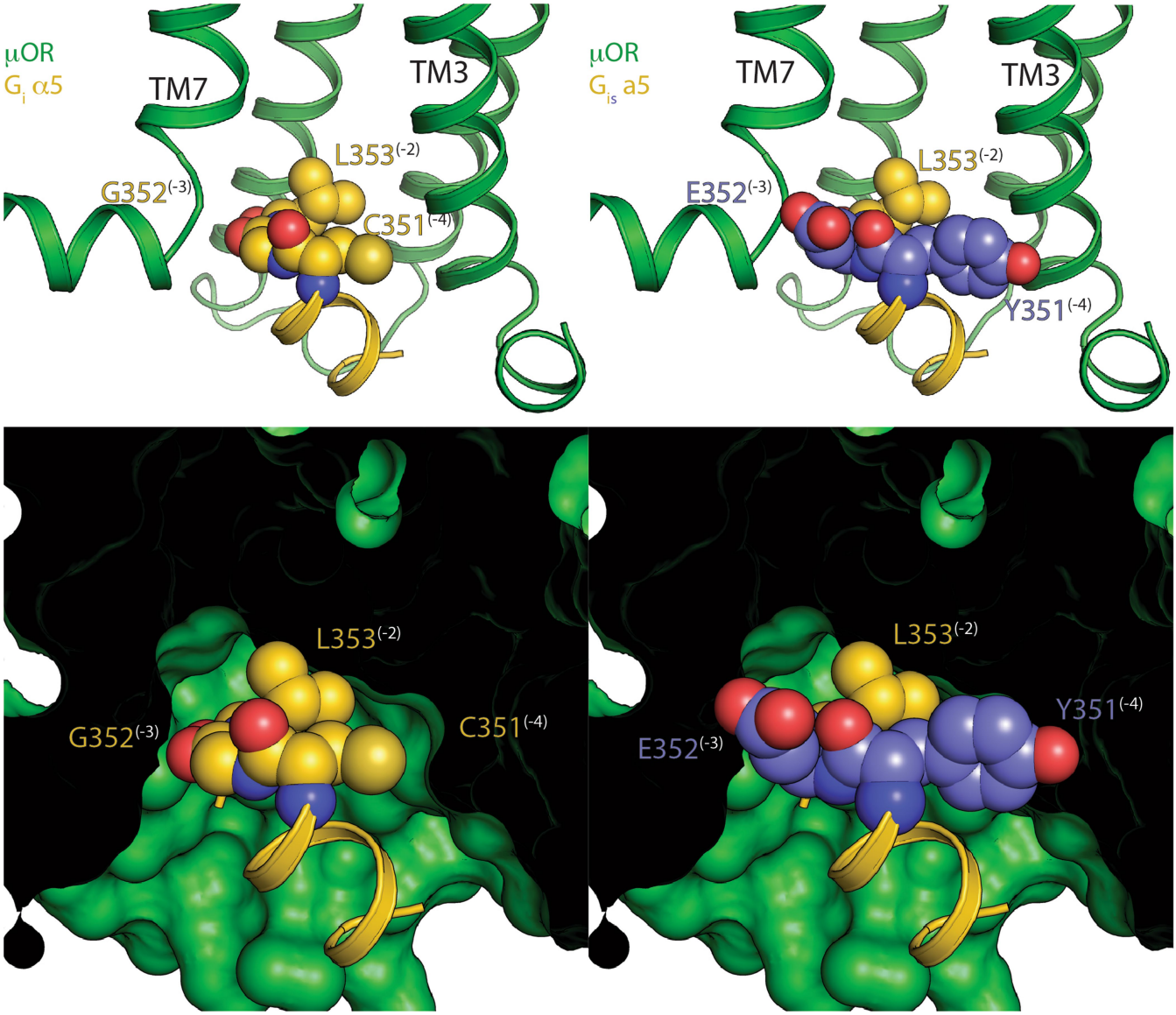
Extended Data Fig. 4 | Selected cryo-EM densities of μOR–G_i complex. **a**–**d**, Cryo-EM density (displayed as mesh) surrounding residues involved in DAMGO binding (**a**), μOR–$G\alpha_i$ interaction around ICL2 (**b**), ICL3 (**c**), and cytoplasmic ends of the μOR transmembrane helices (**d**). These figures accompany the models shown in Figs. 1e, 4b, 5a and 5c, respectively.



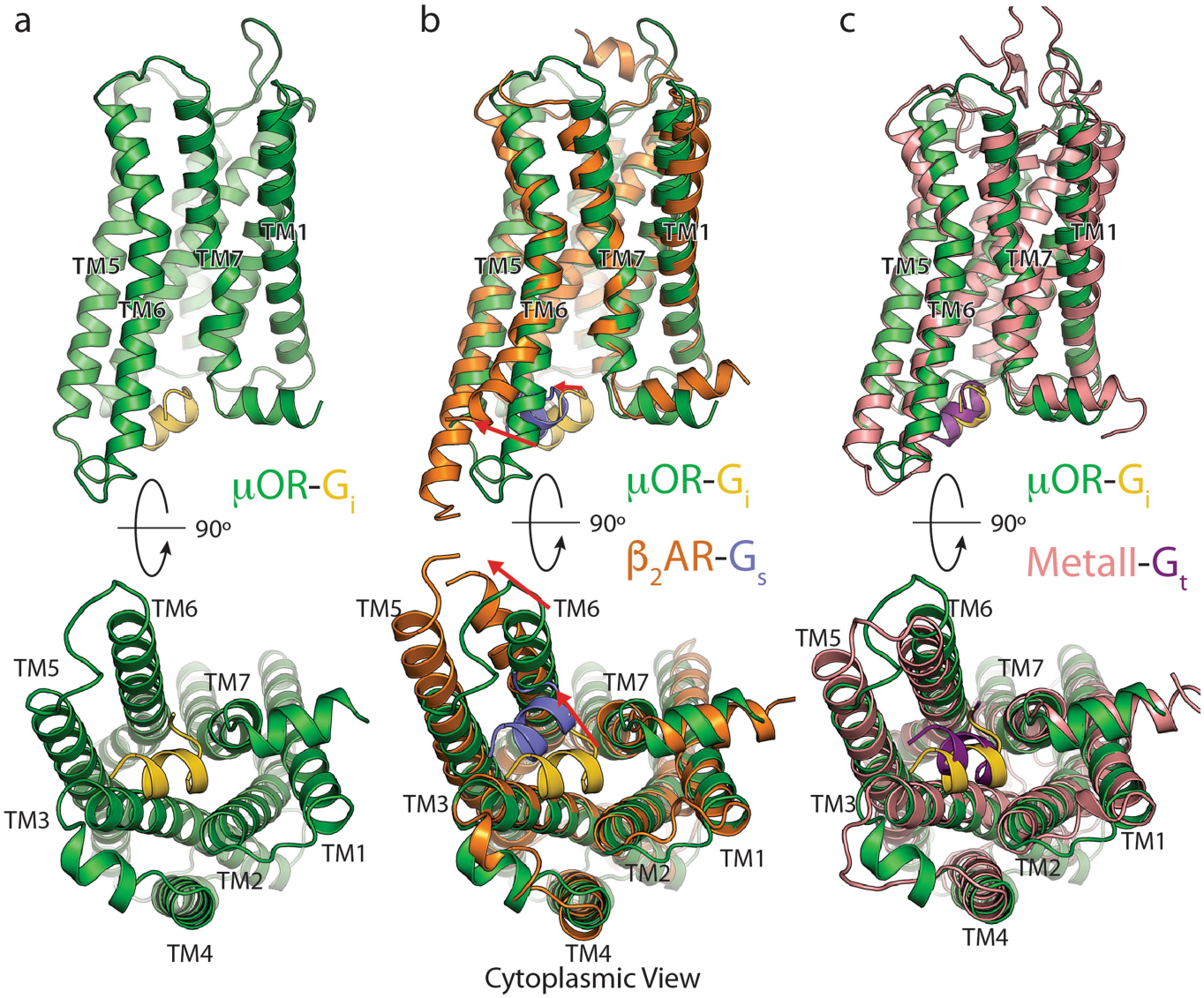
Extended Data Fig. 5 | Stability of DAMGO in molecular dynamics simulations. a, Over the course of molecular dynamics simulations, the positions of the first four residues of DAMGO do not significantly change, while the fifth residue (Gly-ol) shows significant variability in position. Frames from the first and last 100 ns are shown with an intermediate to highlight both the relative stability of the first four amino acids and the flexibility of the fifth. Arrows show the extent of motion in the N- and C-terminal residues over the course of simulation. Cryo-EM density for DAMGO is shown as mesh. **b**, Root mean standard deviations (RMSDs) from the modelled pose of DAMGO to the pose during molecular dynamics simulations. The RMSD calculations include heavy atoms on the peptide backbone. Data from three independent simulations are plotted. The RMSDs for residues 1 to 4 (black) and the C-terminal Gly-ol (blue) are plotted separately to highlight their stability and mobility, respectively.



Extended Data Fig. 6 | Water occupancy in orthosteric binding site. Left, water occupancy in molecular dynamics simulations of DAMGO-bound μOR overlaid with a representative conformation from molecular dynamics simulations. Occupancy relative to bulk solvent is the ratio of the rate at which water is observed in a given volume to the rate at which water is expected to be observed in an equivalent volume in the bulk solvent. For example, blue regions (occupancy ratio = 2) are occupied by water twice as often as an equivalent region in the bulk solvent. Right, crystallographic waters in the BU72-bound μOR binding pocket (PDB code 5C1M). Waters are shown as black spheres, BU72 is shown as yellow sticks, and hydrogen bonds are shown as dashed lines.



Extended Data Fig. 7 | Comparison of the C termini of $G\alpha_s$ and $G\alpha_i$. The C terminus of $G\alpha_s$ is bulkier than that of $G\alpha_i$ owing to substitution of small amino acids C (−4 position) and G (−3 position) in $G\alpha_i$ to Y and E, respectively, in $G\alpha_s$. This leads to steric clashes with TM3 and TM7 of the μOR. Top, ribbon view of μOR (green) with wild-type $G\alpha_i$ (gold, left) and a $G\alpha_{is}$ model (right) created by substituting C and G for Y and E based on the β_2AR–G_s crystal structure. Substituted positions are coloured in light purple. The −4 to −2 positions have their side chains shown as spheres, and the rest are shown as a ribbon. Bottom, space-filling view of the μOR showing the steric clashes that result from these substitutions.



Extended Data Fig. 8 | Comparison of $G\alpha_i$ C-terminal peptide binding modes. a–c, Side (top) and cytoplasmic (bottom) views of the μOR (green) with the last 11 residues of $G\alpha_i$ (gold) alone (**a**), compared to the β_2AR(orange) with the last 11 residues of $G\alpha_s$ (light purple) (PDB code 3SN6) (**b**), or compared to metarhodopsin II (pink) in complex with an 11-residue $G_{transducin}$ (G_t) C-terminal peptide (dark purple) (PDB code 3PQR) (**c**). The μOR–G_i complex aligns best with the metarhodopsin II–G_t complex in terms of both TM6 displacement and position of the α5 peptide.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	μOR-G_i Complex (EMDB-7869) (PDB 6DDF)	μOR-G_i-scFv16 Complex (EMDB-7868) (PDB 6DDE)
Data collection and processing		
Magnification	48,076	48,076
Voltage (kV)	300	300
Electron exposure (e–/Å^2)	40	40
Defocus range (μm)	-0.8 ~ -2.6	-0.8 ~ -2.6
Pixel size (Å)	1.04	1.04
Symmetry imposed	C1	C1
Initial particle images (no.)	893,426	893,426
Final particle images (no.)	359,406	359,406
Map resolution (Å)	3.5 Å	3.6 Å
FSC threshold	(0.143)	(0.143)
Map resolution range (Å)	3.3-4.5	3.3-4.5
Refinement		
Initial model used (PDB code)	5C1M 1GP2	5C1M 1GP2
Model resolution (Å)	3.5	3.6
Model resolution range (Å)	3.3-4.5	3.3-4.5
Map sharpening *B* factor (Å^2)	Pre -90, post -60	Pre -90, post -60
Model composition		
Non-hydrogen atoms	6986	8731
Protein residues	886 residues (6949 atoms)	1119 residues (8694 atoms)
Ligands	1 (37 atoms)	1 (37 atoms)
B factors (Å^2)		
Protein	33.23	69.81
Ligand	31.27	100.66
R.m.s. deviations		
Bond lengths (Å)	0.007	0.010
Bond angles (°)	1.303	1.328
Validation		
MolProbity score	1.93	1.83
Clashscore	7.77	6.97
Poor rotamers (%)	0.72	0.46
Ramachandran plot		
Favored (%)	91.54	93.02
Allowed (%)	8.35	6.89
Disallowed (%)	0.11%	0.09%

Extended Data Table 2 | Sequence alignment of residues that form the interaction interface between μOR and G_i

	Coupling	Branch											
Mu-Opioid Receptor Residue			T103	V169	P172	V173	D177	R179	T180	M255	K271	R277	I278
			2.39	3.54	34.50	34.51	34.55	34.57	4.38	5.61	6.26	6.32	6.33
[Human] 5-HT1A receptor	Gi	α	A	I	P	I	N	R	T	I	A	K	T
[Human] 5-HT1B receptor	Gi	α	A	I	A	V	A	R	T	I	M	K	A
[Human] M2 receptor	Gi	α	N	V	P	L	V	R	T	I	P	K	V
[Human] M4 receptor	Gi	α	N	V	P	L	A	R	T	I	M	K	V
[Human] α2A-adrenoceptor	Gi	α	Q	I	A	I	L	R	T	I	R	R	F
[Human] FPR1	Gi	γ	T	V	P	V	N	R	T	I	-	R	P
[Human] FPR2/ALX	Gi	γ	T	V	P	V	N	R	T	I	-	R	P
[Human] GAL1 receptor	Gi	γ	T	I	S	R	S	R	V	V	-	K	T
[Human] GAL3 receptor	Gi	γ	T	V	P	L	A	R	T	T	R	R	A
[Human] δ Opioid Receptor	Gi	γ	T	V	P	V	D	R	T	M	K	R	I
[Human] κ Opioid Receptor	Gi	γ	T	V	P	V	D	R	T	M	K	R	I
[Human] μ Opioid Receptor	Gi	γ	T	V	P	V	D	R	T	M	K	R	I
[Human] NOP receptor	Gi	γ	T	I	P	I	D	R	T	M	K	R	I
[Human] SST1 receptor	Gi	γ	T	V	P	I	R	R	R	I	R	K	I
[Human] SST2 receptor	Gi	γ	T	V	P	I	K	R	R	I	R	K	V
[Human] SST3 receptor	Gi	γ	T	V	P	T	R	R	T	I	R	R	V
[Human] SST4 receptor	Gi	γ	T	V	P	L	T	R	R	I	R	K	I
[Human] SST5 receptor	Gi	γ	T	V	P	L	R	R	R	I	-	K	V
[Human] CCR1	Gi	γ	T	I	A	V	R	R	T	I	-	K	A
[Human] CCR4	Gi	γ	T	I	A	V	R	R	T	I	-	K	A
[Human] CXCR4	Gi	γ	T	I	A	T	R	R	K	I	-	K	A
[Human] A1 receptor	Gi	α	T	V	P	L	M	V	T	V	Y	K	I
[Human] β1-adrenoceptor	Gs	α	T	I	P	F	S	L	T	V	V	K	A
[Human] β2-adrenoceptor	Gs	α	T	I	P	F	S	L	T	V	F	K	A
[Human] MC1 receptor	Gs	α	M	I	A	L	S	V	T	G	-	K	G
[Human] MC2 receptor	Gs	α	M	I	A	L	S	V	T	K	-	K	G
[Human] MC4 receptor	Gs	α	M	I	A	L	N	M	T	R	-	K	G
[Human] A2A receptor	Gs	α	T	I	P	L	G	V	T	I	T	H	A
[Human] H2 receptor	Gs	α	T	V	P	L	V	V	T	I	A	K	A
[Human] TA1 receptor	Gs	α	T	V	P	L	A	M	N	I	S	K	A
[Human] RXFP1	Gs	δ	Y	I	P	F	R	-	G	M	Q	I	L
[Human] RXFP2	Gs	δ	H	I	P	F	R	-	G	M	C	A	V
[Human] V2 receptor	Gs	β	I	I	P	M	R	G	S	I	V	K	T
[Human] 5-HT2A receptor	Gq	α	T	I	P	I	R	N	S	T	S	K	A
[Human] 5-HT2B receptor	Gq	α	T	I	P	I	Q	N	S	T	T	R	A
[Human] M1 receptor	Gq	α	N	V	P	L	A	R	T	I	S	K	A
[Human] M3 receptor	Gq	α	N	I	P	L	A	R	T	I	S	K	A
[Human] M5 receptor	Gq	α	N	I	P	L	A	R	T	I	V	K	A
[Human] α1A-adrenoceptor	Gq	α	T	V	P	L	T	V	T	V	K	K	A
[Human] GAL2 receptor	Gq	γ	T	I	P	L	E	R	T	T	A	K	V
[Human] OX1 receptor	Gq	β	T	I	P	L	-	-	T	I	Q	K	T
[Human] OX2 receptor	Gq	β	T	I	P	L	-	-	T	I	Q	K	T
[Human] NK1 receptor	Gq	β	T	I	-	-	-	-	S	V	Q	K	V

Receptors from different branches of the GPCR family with different coupling specificities were selected for analysis. Sequences and alignment were performed using GPCRDB (http://www.gpcrdb.org)

ARTICLE

https://doi.org/10.1038/s41586-018-0215-y

Cryo–EM structure of human rhodopsin bound to an inhibitory G protein

Yanyong Kang[1,12], Oleg Kuybeda[2,12], Parker W. de Waal[1,12], Somnath Mukherjee[3], Ned Van Eps[4], Przemyslaw Dutka[3,5], X. Edward Zhou[1], Alberto Bartesaghi[2], Satchal Erramilli[3], Takefumi Morizumi[4], Xin Gu[1], Yanting Yin[1], Ping Liu[6,7], Yi Jiang[7], Xing Meng[8], Gongpu Zhao[8], Karsten Melcher[1], Oliver P. Ernst[4,9], Anthony A. Kossiakoff[3,10]*, Sriram Subramaniam[2,11]* & H. Eric Xu[1,7]*

G-protein-coupled receptors comprise the largest family of mammalian transmembrane receptors. They mediate numerous cellular pathways by coupling with downstream signalling transducers, including the hetrotrimeric G proteins G_s (stimulatory) and G_i (inhibitory) and several arrestin proteins. The structural mechanisms that define how G-protein-coupled receptors selectively couple to a specific type of G protein or arrestin remain unknown. Here, using cryo-electron microscopy, we show that the major interactions between activated rhodopsin and G_i are mediated by the C-terminal helix of the G_i α-subunit, which is wedged into the cytoplasmic cavity of the transmembrane helix bundle and directly contacts the amino terminus of helix 8 of rhodopsin. Structural comparisons of inactive, G_i-bound and arrestin-bound forms of rhodopsin with inactive and G_s-bound forms of the β_2-adrenergic receptor provide a foundation to understand the unique structural signatures that are associated with the recognition of G_s, G_i and arrestin by activated G-protein-coupled receptors.

The selective coupling of cell-surface receptors with specific intracellular effector proteins is a fundamental step in transmembrane signalling. G-protein-coupled receptors (GPCRs) constitute the largest protein family of transmembrane receptors with more than 800 members in humans[1]. These receptors signal primarily through intracellular G proteins and arrestin proteins. Compared to the vast diversity of the GPCR family members and their physiological functions, the number of intracellular signalling transducers are much limited, with only four major types of G protein[2] (Fig. 1a) and two major types of arrestin[3]. These distinct signalling transducers regulate the generation of a variety of secondary messengers and activate various downstream kinases, which lead to diverse cellular signalling pathways and physiological consequences. Thus, the selective coupling of a specific transducer protein by a GPCR is crucial for cellular signalling and responses to extracellular stimuli.

Our understanding of GPCR signalling has been greatly enhanced by the remarkable progress in GPCR structural biology, including the determination of over 170 structures from 40 unique GPCR members in active or inactive states and in complex with G_s or arrestin[4]. In the inactive states, the cytoplasmic end of the seven-transmembrane-helix domain (TMD) is closed[5–7], thus preventing GPCRs from effective coupling with cellular signalling transducers. Agonist binding induces conformational changes in the TMD, including an outward movement at the cytoplasmic end of transmembrane helix 6 (TM6)[8–13], which opens a cavity in the TMD bundle for interaction with one or more specific transducer proteins. This mechanism is particularly highlighted by the notable outward movement (14–18 Å) at the cytoplasmic side of TM6 in the G_s-bound β_2-adrenergic receptor (β_2AR) structure[14] and the recent cryo-electron microscopy (cryo-EM) structures of G_s-bound class B GPCRs[15,16]. Combination of the structural observations with sequence analyses has revealed a barcode system in G proteins for GPCR–G-protein binding selectivity[17]. However, less is known about the mechanisms determining selectivity on the receptor, in particular the subset of G_i-coupled receptors, despite the fact that they constitute the largest fraction of GPCR proteins[17].

Rhodopsin is a prototypical GPCR that plays a crucial role in light perception and has served as a model system for studying GPCR signalling[18–20]. In the pre-illuminated state, rhodopsin adopts an inactive conformation stabilized by the inverse agonist, 11-*cis*-retinal. Light induces isomerization of the retinal ligand to the all-*trans*-retinal (ATR) configuration, which activates the receptor. The activated rhodopsin is coupled to the G protein transducin (G_t) (Fig. 1a), which is a G_i homologue[21,22], to initiate the light-sensing signalling pathways. To understand the structural basis for the selective coupling of G_i by a GPCR, we determined the structure of the rhodopsin–G_i complex at a near atomic resolution using cryo-EM.

Cryo–EM of the rhodopsin–G_i complex

To prepare a stable rhodopsin–G_i complex for structural studies, we took advantage of a constitutively active form of rhodopsin (termed 4 M, containing mutations $N2^{Nterm}C$, $E113^{3.28}Q$, $M257^{6.40}Y$, $N282^{ECL3}C$; superscripts denote Ballesteros–Weinstein numbering), that was previously used to determine the structure of the rhodopsin–arrestin complex[23,24]. In cell-based assays, the constitutively active rhodopsin activates the G_i coupling pathway more strongly than the wild-type receptor (Fig. 1b). Importantly, wild-type and constitutively active rhodopsin do not activate the G_s pathway regardless of the presence of ATR (Extended Data Fig. 1c). Because previous studies indicated that the

[1]Center for Cancer and Cell Biology, Innovation and Integration Program, Van Andel Research Institute, Grand Rapids, MI, USA. [2]Cancer Research Technology Program, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. [3]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois, USA. [4]Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada. [5]Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Krakow, Poland. [6]University of Chinese Academy of Sciences, Beijing, China. [7]Key Laboratory of Receptor Research, VARI-SIMM Center, Center for Structure and Function of Drug Targets, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China. [8]David Van Andel Advanced Cryo-Electron Microscopy Suite, Van Andel Research Institute, Grand Rapids, MI, USA. [9]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [10]Institute for Biophysical Dynamics, University of Chicago, Chicago, IL, USA. [11]Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD, USA. [12]These authors contributed equally: Yanyong Kang, Oleg Kuybeda, Parker W. de Waal. *e-mail: koss@bsd.uchicago.edu; ss1@nih.gov; eric.xu@vai.org

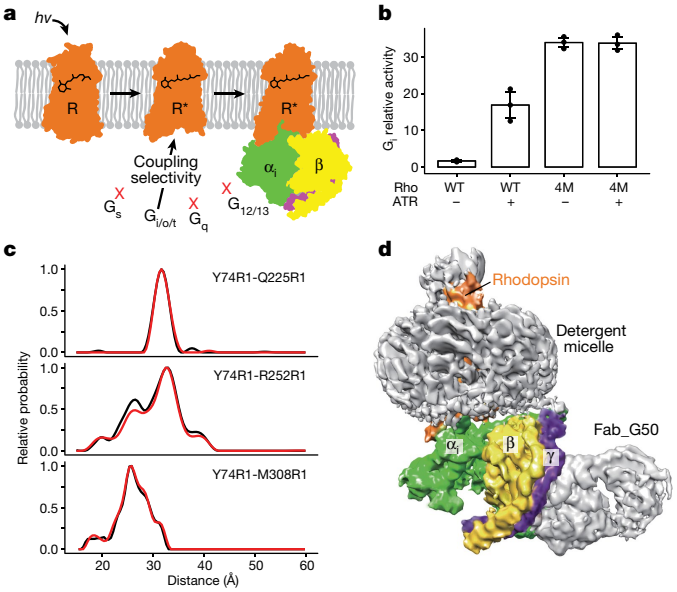


Fig. 1 | Assembly of the rhodopsin–G_i protein complex. a, Schematic illustration of G-protein-mediated GPCR signalling by the four types of G protein. Light-activated rhodopsin is specifically coupled to the $G_{i/o/t}$ subtype. **b**, G_i signalling activated by the 4M mutant and wild-type (WT) rhodopsin as measured by a serum response element (SRE)-driven luciferase reporter assay ($n = 3$). Data are mean ± s.d. **c**, Experimental DEER distance distributions of rhodopsin in the presence (red line) or absence (black line) of Fab. Y74R1-Q225R1, Y74R1-R252R1 and Y74R1-M308R1 denote R1 nitroxide pairs. **d**, Iso-surface rendering of the cryo-EM density map for the rhodopsin–G_i–Fab complex.

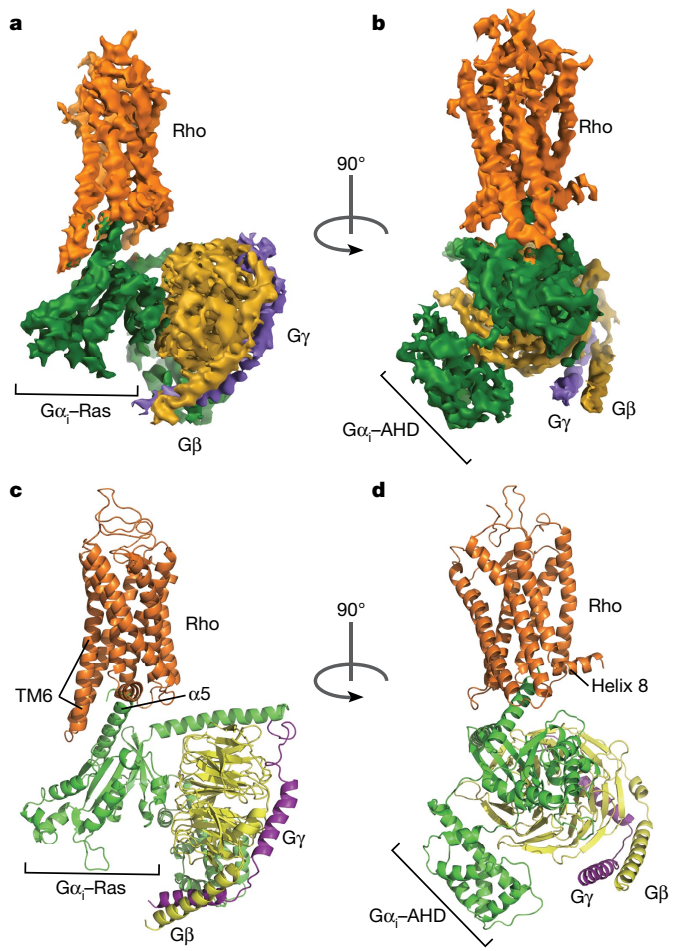


Fig. 2 | The cryo-EM structure of the rhodopsin–G_i complex. **a**, **b**, Orthogonal views of the cryo-EM density map of the rhodopsin–G_i complex, coloured by subunits. **c**, **d**, Ribbon diagram representation of the structure of the rhodopsin–G_i complex.

active rhodopsin–G_i complex is more stable than the active rhodopsin–G_t complex[25], we focused on obtaining a stable complex of activated rhodopsin bound to a dominant-negative mutant form of G_i (Extended Data Fig. 1a, b), which was used to promote the nucleotide-free form of G_i[26]. To further stabilize the rhodopsin–G_i complex, we screened a panel of antibodies against the dominant-negative G_i from a phage display Fab library by negative-stain electron microscopy to identify one Fab fragment, termed Fab_G50, which stabilized the rhodopsin–G_i complex. The conformational state of rhodopsin in complex with G_i and Fab was surveyed by high-resolution distance mapping with site-directed spin labelling and double electron–electron resonance (DEER) spectroscopy in detergent micelles (Fig. 1c and Extended Data Fig. 1d–f). Almost identical distances were measured when rhodopsin was bound to G_i in the presence or absence of Fab_G50, suggesting that the Fab fragment does not affect the conformation of rhodopsin. In addition, the distance distributions of TM6 and TM7 relative to TM2 are nearly identical between G_i-bound and G_t-bound rhodopsin[27,28] (Extended Data Fig. 1e).

The cryo-EM structure of the rhodopsin–G_i–Fab complex was determined at a nominal global resolution of 4.5 Å (Fig. 1d and Extended Data Fig. 2a–f), which reveals a well-defined density map for the rhodopsin seven-transmembrane bundle, the $G\alpha_i$ Ras-like domain, and the Gβ and Gγ subunits (Fig. 2a, b and Extended Data Fig. 3a–g). The position of the α-helical domain (AHD) of $G\alpha_i$ is well defined owing to the direct stabilization of this domain by the Fab fragment, which simultaneously interacts with the $G\alpha_i$ AHD and the Gβ subunit (Fig. 1d). The binding site of the Fab fragment is far away from the rhodopsin–G_i interface, consistent with the fact that Fab binding did not affect the conformation of rhodopsin.

The rhodopsin–G_i interface

The structure of the rhodopsin–G_i complex (Fig. 2a, b and Extended Data Table 1) reveals that interactions between rhodopsin and G_i are exclusively mediated through the $G\alpha_i$ subunit, and that there is no contact between rhodopsin and the Gβγ subunits. The most important interface between rhodopsin and $G\alpha_i$ is formed by the last 11 residues of the C-terminal α5-helix of $G\alpha_i$ (Fig. 2c, d), consistent with its known crucial role in receptor binding[29–35]. In the complex, the C-terminal 11 residues (residues 344–354) of $G\alpha_i$ adopt a straight amphipathic helix (α5-helix) to residue C351, followed by a three-residue loop (residue 352-GLF-354) (Fig. 3a, b).

The negatively charged dipole of the C terminus of the α5-helix forms an electrostatic interaction with the positively charged dipole of the N terminus of helix 8 of rhodopsin, with the carbonyl group of G352 forming the capping interaction with the N-terminal amine groups of helix 8 of rhodopsin (Fig. 3c). The cytoplasmic side of the rhodopsin transmembrane bundle is highly positively charged and this is known to be a common feature of GPCR structures[23]. Sequence alignment of all $G\alpha_i$ and $G\alpha_t$ subtypes reveals a conserved sequence pattern (Fig. 3d), including D350 and G352, which is at the terminus of the α5-helix and allows the uncapped carbonyl group of α5-helix to form charge interactions with helix 8 of rhodopsin (Fig. 3b). The G352C-mutated $G\alpha_i$ did not crosslink with rhodopsin with cysteine mutations around the transmembrane bundle (Extended Data Fig. 4a, b), consistent with the finding that the G352C mutation in G_i disrupts its binding capability to rhodopsin.

Beside the charge interactions, the two large hydrophobic side chains L353 and L348 are directed towards the hydrophobic pocket of rhodopsin formed by TM3, TM5, TM6 and TM7 (Fig. 3a, b). Replacement of these two residues by alanine, together with G352A, has marked effects on G_i binding to rhodopsin[36]. Residues I344, K345 and C351 of $G\alpha_i$ are also within packing distance with TM3, TM5 and TM6 (Fig. 3a, b), and alanine mutations in these residues also show reduced G_i binding

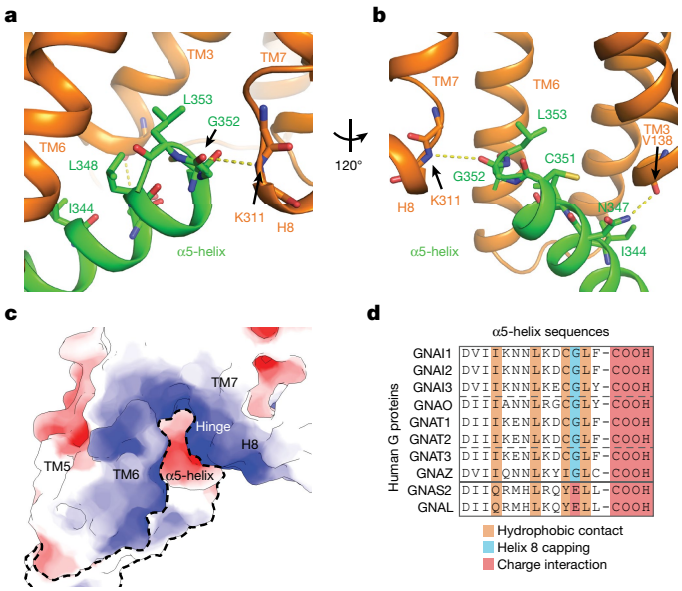


Fig. 3 | The rhodopsin–G_i interface. a, **b**, Two views of the binding interface between the α5-helix of the $G\alpha_i$ Ras-like domain and the TMD cavity of rhodopsin. **c**, Rendering of electrostatic surfaces involved in interaction of rhodopsin and $G\alpha_i$, with blue for positively charged regions and red for negatively charged regions. **d**, Sequence alignment of the last 11 residues of the α5-helix from different G proteins, with key residues in receptor binding highlighted by colour shading.

to rhodopsin[36]. Thus, our structural studies provide a structural explanation to rationalize the extensive mutagenesis studies on the effect of the last 11 residues of the C terminus of $G\alpha_i$[36].

The second and less extensive interface between rhodopsin and G_i is mediated through the N-terminal helix (αN-helix) of $G\alpha_i$, where its residues (E28 and R32) are in proximity to interact with the intracellular loop (ICL2) between TM3 and TM4 (Extended Data Fig. 5c). In contrast to a short α-helix formed by the ICL2 loop in the arrestin-bound rhodopsin structure[23], ICL2 in the G_i-bound complex adopts an extended loop. The presence of the interface between ICL2 and αN-helix was confirmed by site-specific disulfide crosslinking from cysteine mutations in residue E28 in $G\alpha_i$ to residues $N145^{ICL2}$ and $F146^{ICL2}$ in ICL2 of rhodopsin (Extended Data Fig. 4c, d), in agreement with previous crosslinking results between $G\alpha_t$ and light-activated rhodopsin by a chemically activated crosslinking reagent[37]. Furthermore, R32A in $G\alpha_i$ is a rare mutation that increases the binding of $G\alpha_i$ to rhodopsin[36]. On the basis of the structure, the large side chain of R32 in $G\alpha_i$ is in the position that could interfere with the close interactions between the αN-helix in $G\alpha_i$ and the ICL2 of rhodopsin and the removal of the large side chain in R32A mutations could thus enhance rhodopsin–Gi interactions.

Crystal structures of rhodopsin have been determined in several different states, including the inactive/inverse agonist-bound state[5] and the active arrestin-bound or GαCT-bound state[10,12] (GαCT is a high-affinity peptide from the C terminus of $G\alpha_t$). Comparisons of these structures reveal intriguing conformational differences in rhodopsin between each state (Fig. 4a). Compared to the inactive state, the G_i-bound rhodopsin structure has an extended TM5 and outward movements in TM6, TM1 and TM4 at the cytoplasmic side (Fig. 4a, Supplementary Video 1), which results in an elastic pocket for binding the G protein. The overall structure of rhodopsin bound to G_i is similar to that bound to the GαCT, consistent with the equivalence of rhodopsin coupling with G_i and G_t (Fig. 4b). The conformation of the C-terminal 11 residues overlaps well with the structure of GαCT. However, the position of the GαCT is shifted a half-helical turn (approximately 2.2 Å) deeper into the rhodopsin helical bundle owing to the lack of constrain on the GαCT peptide (Fig. 4b).

Structural comparison of rhodopsin in the G_i-bound and arrestin-bound states also reveals notable differences in the receptor conformation, including TM1, TM4, TM6 and helix 8 (Fig. 4a, Supplementary Video 2), as well as ICL2, which adopts a short α-helix in arrestin-bound rhodopsin but exists as an extended loop in the G_i-bound state (Fig. 4a). These conformational differences may represent the unique structural signatures of rhodopsin to distinguish between G protein and arrestin. In addition, other factors such as phosphorylation at the rhodopsin C-terminal tail may further enhance the specific interactions of arrestin with the activated receptor[24].

Structural basis of G_i and G_s selectivity

Rhodopsin and β_2AR are the two best studied GPCRs that are coupled to G_i and G_s, respectively. Structural comparisons between these two receptors reveal the basis for selective coupling of G_i and G_s by a GPCR. In the inactive state, both receptors adopt a very similar closed conformation, in which the cytoplasmic ends of TM6, TM1, TM2, TM4 and the turn between TM7 and helix 8 are in nearly identical positions to each other (Fig. 5a, b). In the active state, the cytoplasmic ends of TM2 and TM4 of both receptors remain in identical position, but the cytoplasmic ends of TM6, TM1 and the turn between TM7 and helix 8 differ considerably between the two receptors (Fig. 5c, d, Supplementary Video 3). In particular, the cytoplasmic end of TM6 in rhodopsin is shifted inward by approximately 8 Å relative to that of TM6 seen in the G_s-bound β_2AR complex or in the G_s-bound class B GPCR structures (Extended Data Fig. 5a–d). Correspondingly, the α5-helix of $G\alpha_i$ is rotated 20° away from TM6 towards helix 8 and forms a capping interaction with the N-terminal dipole of helix 8 (Fig. 3a, b).

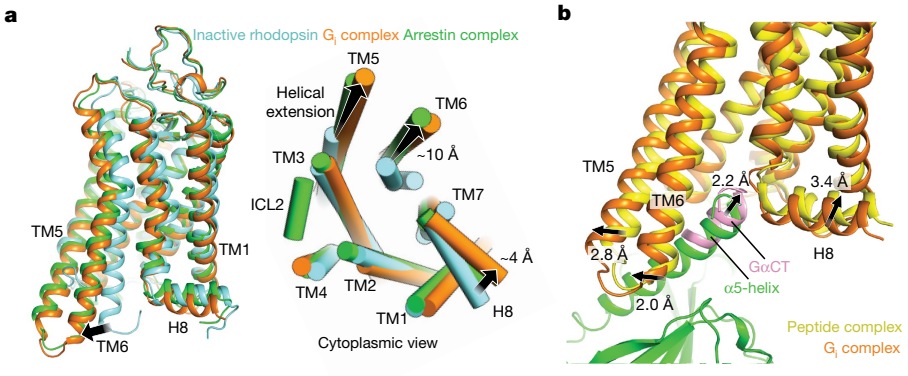


Fig. 4 | Structural comparison of G_i-bound rhodopsin with inactive rhodopsin, arrestin-bound rhodopsin, and GαCT-bound rhodopsin. a, Side and cytoplasmic views of the G_i-bound transmembrane bundle (orange) in superposition to the inactive rhodopsin (PDB code 1U19, cyan) and arrestin-bound rhodopsin (PDB code 4ZWJ, green). **b**, Superposition of G_i-bound rhodopsin (orange) with GαCT-bound rhodopsin (yellow). Differences in transmembrane domains at the cytoplasmic faces are highlighted.

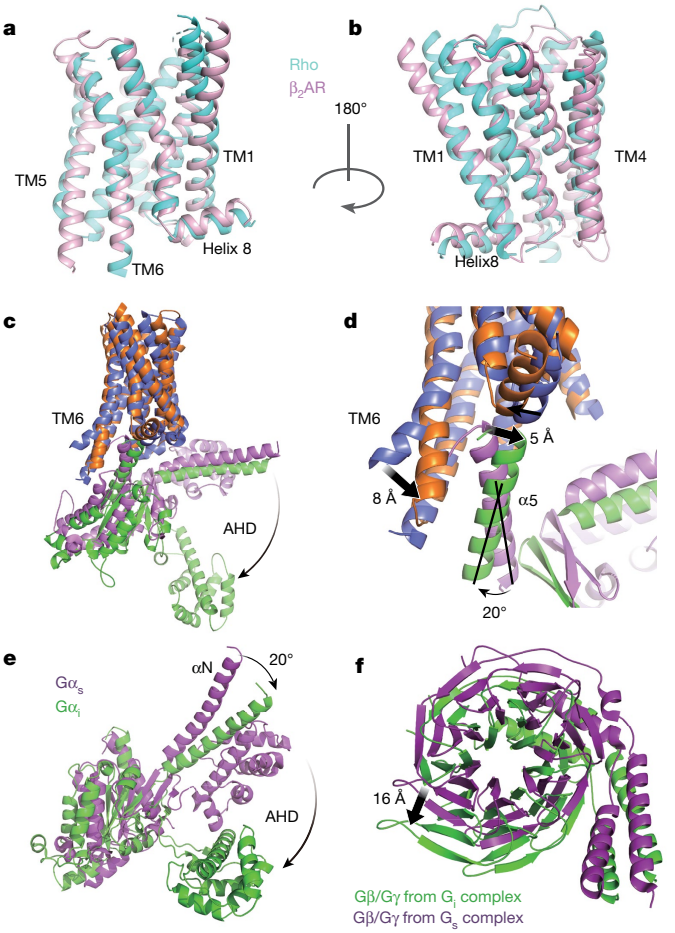


Fig. 5 | Structural comparison of G_i-bound rhodopsin with G_s-bound β_2AR. a, **b**, Side view of the inactive rhodopsin structure (PDB code 1U19, cyan) superposed with inactive β_2AR (PDB code 2RH1, pink). **c**, **d**, Side and cytoplasmic views of G_i-bound rhodopsin compared to G_s-bound β_2AR (PDB code 3SN6, blue). Notable structural changes are seen for the intracellular domains of TM6 with a difference of approximately 8 Å at the cytoplasmic end of the helix. The α5 helix of $G\alpha_i$ (green) is rotated 20° away from TM6 compared to that of $G\alpha_s$ (purple). As indicated, there are differences in the locations of the α-helical domains (AHD) of these two G proteins. **e**, **f**, Illustration of the 20° rotation of $G\alpha_i$ (**e**) and the 16 Å shift in $\beta\gamma$ as compared to the structure of G_s (**f**).

Compared to the straight α5-helix in $G\alpha_i$, the α5-helix in $G\alpha_s$ is slightly kinked (Fig. 5d), which allows the C terminus of this helix to orient away from helix 8 towards TM6 of rhodopsin. Accompanying the 20° rotation of the α5-helix between $G\alpha_i$ and $G\alpha_s$ is a rigid body rotation of the G_i heterotrimer, which displays rearrangements up to 16 Å in the areas of the $G\alpha_i$ N-terminal end and Gγ subunits (Fig. 5e, f). The observed differences between binding of G_i and G_s to their receptors are in good agreement with a model of the rhodopsin–Gi complex that resulted from experimental DEER distance mapping[28]. On the basis of these structural observations, we reason that the conformational difference in the TM6 between rhodopsin and β_2AR is one of the major determinants for the coupling specificity of G_i and G_s in the two receptors. This movement of TM6 upon activation is a general theme in seven-helix transmembrane receptors, and was originally observed in cryo-EM studies of light-induced conformational changes in the proton pump bacteriorhodopsin[38,39].

An additional cause of the coupling specificity of G_i and G_s between rhodopsin and β_2AR may reside in the regions of the TM5–α4-helix interface (Extended Data Fig. 5e, f). Superposition of G_s and G_i in the two receptor complexes reveals a steric collision between the extended TM5 of β_2AR and α4-helix of $G\alpha_i$, consistent with the involvement of this region in receptor coupling specificity[40,41].

Structural comparison of our nucleotide-free G_i complex with the GDP-bound inactive G_i complex also reveals an interesting mechanism of rhodopsin-mediated G_i activation (Extended Data Fig. 6a). Upon binding to rhodopsin, the α5-helix was rotated by 90° and extended by two additional helical turns into the rhodopsin cytoplasmic pocket (Extended Data Fig. 6b), which resulted in conformational changes in the loop between β6 and α5-helix, a key structural element of the GDP-binding pocket. This conformational change leads to a movement of α1-helix into the GDP-binding pocket, thus expelling the GDP from its pocket (Extended Data Fig. 6c, d). The release of GDP from $G\alpha_i$ induces separation of the $G\alpha_i$ Ras domain from its AHD, the conformation of which was captured by the Fab fragment as observed in our structure. The mechanism of rhodopsin-mediated G_i activation is similar to that of β_2AR-mediated G_s activation[14,42,43], suggesting a common mechanism of GPCR-mediated G protein activation.

To gain further insight into differential conformational dynamics of TM6 between active rhodopsin and β_2AR, we performed all-atom mollified adaptive biasing potential (mABP) simulations of the receptors with the fABMACS software package to accelerate sampling[44] (Extended Data Table 2). By calculating the free energy landscape of TM6 movement relative to the receptor transmembrane bundle (Extended Data Fig. 7a, b), these biased simulations revealed that the dynamic range of TM6 of rhodopsin is considerably less than that of β_2AR (Fig. 6a–c). Furthermore, the TM6 of rhodopsin was unable to swing outwards to sample conformations comparable to that of β_2AR, resulting in steric clashes between rhodopsin TM6 and G_s, consistent with the inability of rhodopsin to couple with G_s (Fig. 6c and Extended Data Fig. 1b). To extend this mechanism to other class A GPCRs, we performed similar simulations for the μ-opioid receptor 1 (μOR1)[45], a G_i-coupled receptor, and the adenosine A_{2A} receptor (A_{2A}R)[20], a G_s-coupled receptor. These simulations also revealed that TM6 of μOR1 remained in a more closed conformation, whereas TM6 of A_{2A}R favoured an outward conformation (Fig. 6a–c and Extended Data Fig. 7c, d). We note however that both G_s coupler simulations were performed in the absence of a complete ICL3, whereas both G_i couplers had their ICL3 intact. Additional simulations with ICL3-truncated G_i couplers revealed no notable differences compared to the full-length receptor (Fig. 6b), indicating that their ICL3 was not sufficient to restrict the outward movement of TM6. Together, these results suggest that conformational differences in TM6 may represent the basis for stratification of GPCRs into distinct conformational groups with respect to the coupling specificity for G_i or G_s by GPCRs[46] (Fig. 6d).

TM6 sequence motif for G_i selectivity

To investigate the basis for the different TM6 dynamics between G_i and G_s coupling GPCRs, we performed sequence analysis on TM6, which reveals that G_i- and G_s-coupled receptors exhibit distinct, inversely related enrichment patterns for polar and hydrophobic residues at the membrane interface (Extended Data Fig. 8a–c). For G_i-coupled receptors, an enrichment of polar, often positively charged, residues at TM6.31/34/35 could act to stabilize the receptor within the charged lipid head groups thus preventing outward movement of TM6 (Extended Data Fig. 8c). By contrast, an enrichment of hydrophobic residues found in G_s coupling receptors would promote an outward swing of TM6 favouring interactions with the hydrophobic lipid tails. In addition, we found another potential selectivity filter at TM6.36 where G_i and G_s couplers exhibit differential enrichment of polar/hydrophobic residues. Comparison of our G_i-bound rhodopsin to G_s-bound β_2AR and miniG$_s$-bound A_{2A}R shows that M253$^{6.36}$ of rhodopsin forms a hydrophobic interaction with L353 of $G\alpha_i$, whereas T274$^{6.36}$ of β_2AR and S234$^{6.36}$ of A_{2A}R do not appear to interact with the G_s or miniG$_s$ (Extended Data Fig. 8d–f). Instead these polar residues in G_s-coupled receptors are enriched near an extra kink specifically in the TM6 helix of G_s-coupled receptors. We questioned whether they may have a role to destabilize TM6 and promote its outward kink. Indeed, throughout simulations, both S234 and T274 form extensive hydrogen bonding with the kinked TM6 backbone of β_2AR and A_{2A}R (Extended Data

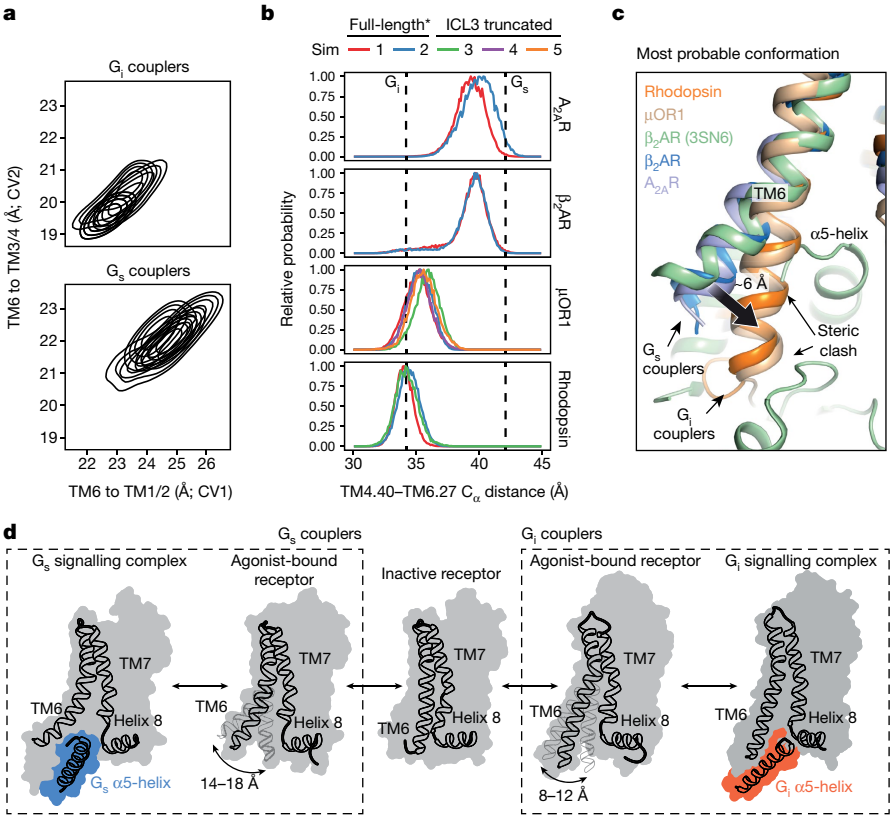


Fig. 6 | TM6 dynamics of G_i- and G_s-coupled receptors. a, G_i-coupled receptors exhibit a markedly constrained range of motion compared to G_s-coupled receptors. Comparison of overlapped free energy landscapes truncated at 10 kJ mol^{-1} for G_i- and G_s-coupling GPCRs plotted with their collective variables used for mABP simulations. CV1 and CV2, collective variables 1 and 2, respectively. **b**, Weighted TM4.40 to TM6.27 C_α distance distributions for full-length (simulations 1 and 2) and ICL3-truncated receptor simulations (simulations 3–5). An asterisk on full-length indicates that ICL3 was absent in both G_s couplers. Reference distances for G_i- and G_s-bound states (shown as black vertical dashed lines) are taken from the structure reported here and the β_2AR–G_s protein complex (PDB code 3SN6). For G_i-coupled receptors with and without a complete ICL3, the outward movement of TM6 is approximately 6 Å less than that of G_s couplers. **c**, Representative snapshots of the most probable TM6 position taken from the first simulation replicate, overlapped with the β_2AR–G_s crystal structure. Steric clashed can be seen between the TM6 of both G_i couplers and G_s. **d**, Schematic depicting alternative TM6 conformational states as the structural determinants for selective coupling of G_i and G_s. TM6 distance ranges were calculated using the structure of inactive rhodopsin (PDB code 1F88) as a fixed reference point.

Fig. 8g). We speculate that the polar/non-polar residue distribution in TM6 may contribute to differential conformational dynamics of TM6 between G_i- and G_s-coupled receptors.

In summary, our results show that the conformational change, especially the outward movement of TM6, is less pronounced for G_i-bound rhodopsin than the corresponding movement in the G_s-bound β_2AR structure. This conformational difference seems to be the key determinant for the different docking of the C-terminal α5-helix between $G\alpha_i$ and $G\alpha_s$ into the receptor transmembrane bundle. Energy landscape analysis of a different set of G_i- and G_s-coupled receptors also revealed that they have similar profiles to those of rhodopsin and β_2AR, respectively, with respect to the position of TM6. Sequence analysis reveals that the differential swing of TM6 between G_i- and G_s-coupled receptors could be attributed to different polar/nonpolar residue distribution in the TM6. Together, these data suggest that the basis for G_i- and G_s-coupling selectivity observed in rhodopsin and β_2AR is a general theme for GPCRs to distinguish between G_i and G_s receptors. G_i-coupled receptors represent the largest subgroup of receptors in the GPCR superfamily. On the basis of the common features of the positively charged transmembrane bundle of GPCRs and the conserved sequences of the α5-helix in $G\alpha_i$, we expect that the structure of the rhodopsin–G_i complex will serve as a model for understanding signalling of other G_i-coupled receptors.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0215-y.

Received: 12 December 2017; Accepted: 2 May 2018;
Published online: 13 June 2018

1. Fredriksson, R., Lagerström, M. C., Lundin, L. G. & Schiöth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **63**, 1256–1272 (2003).
2. Neves, S. R., Ram, P. T. & Iyengar, R. G protein pathways. *Science* **296**, 1636–1639 (2002).
3. Gurevich, E. V. & Gurevich, V. V. Arrestins: ubiquitous regulators of cellular signaling pathways. *Genome Biol.* **7**, 236 (2006).
4. Zhou, X. E., Melcher, K. & Xu, H. E. Understanding the GPCR biased signaling through G protein and arrestin complex structures. *Curr. Opin. Struct. Biol.* **45**, 150–159 (2017).
5. Palczewski, K. et al. Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289**, 739–745 (2000).
6. Cherezov, V. et al. High-resolution crystal structure of an engineered human β_2-adrenergic G protein-coupled receptor. *Science* **318**, 1258–1265 (2007).
7. Rosenbaum, D. M. et al. GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* **318**, 1266–1273 (2007).
8. Xu, F. et al. Structure of an agonist-bound human A_{2A} adenosine receptor. *Science* **332**, 322–327 (2011).
9. Lebon, G. et al. Agonist-bound adenosine A_{2A} receptor structures reveal common features of GPCR activation. *Nature* **474**, 521–525 (2011).
10. Choe, H. W. et al. Crystal structure of metarhodopsin II. *Nature* **471**, 651–655 (2011).
11. Wang, C. et al. Structural basis for molecular recognition at serotonin receptors. *Science* **340**, 610–614 (2013).
12. Standfuss, J. et al. The structural basis of agonist-induced activation in constitutively active rhodopsin. *Nature* **471**, 656–660 (2011).

13. Rosenbaum, D. M. et al. Structure and function of an irreversible agonist–β_2 adrenoceptor complex. *Nature* **469**, 236–240 (2011).
14. Rasmussen, S. G. et al. Crystal structure of the β_2 adrenergic receptor–G_s protein complex. *Nature* **477**, 549–555 (2011).
15. Zhang, Y. et al. Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature* **546**, 248–253 (2017).
16. Liang, Y. L. et al. Phase-plate cryo-EM structure of a class B GPCR-G-protein complex. *Nature* **546**, 118–123 (2017).
17. Flock, T. et al. Selectivity determinants of GPCR–G-protein binding. *Nature* **545**, 317–322 (2017).
18. Palczewski, K. G protein-coupled receptor rhodopsin. *Annu. Rev. Biochem.* **75**, 743–767 (2006).
19. Zhou, X. E., Melcher, K. & Xu, H. E. Structure and activation of rhodopsin. *Acta Pharmacol. Sin.* **33**, 291–299 (2012).
20. Hamm, H. E. How activated receptors couple to G proteins. *Proc. Natl Acad. Sci. USA* **98**, 4819–4821 (2001).
21. Van Meurs, K. P. et al. Deduced amino acid sequence of bovine retinal $G_{o\alpha}$: similarities to other guanine nucleotide-binding proteins. *Proc. Natl Acad. Sci. USA* **84**, 3107–3111 (1987).
22. Lerea, C. L., Somers, D. E., Hurley, J. B., Klock, I. B. & Bunt-Milam, A. H. Identification of specific transducin α subunits in retinal rod and cone photoreceptors. *Science* **234**, 77–80 (1986).
23. Kang, Y. et al. Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. *Nature* **523**,561–567 (2015).
24. Zhou, X. E. et al. Identification of phosphorylation codes for arrestin recruitment by G protein-coupled receptors. *Cell* **170**, 457–469 (2017).
25. Maeda, S. et al. Crystallization scale preparation of a stable GPCR signaling complex between constitutively active rhodopsin and G-protein. *PLoS One* **9**, e98714 (2014).
26. Liu, P. et al. The structural basis of the dominant negative phenotype of the $G\alpha_{i1}\beta_1\gamma_2$ G203A/A326S heterotrimer. *Acta Pharmacol. Sin.* **37**, 1259–1272 (2016).
27. Van Eps, N. et al. Conformational equilibria of light-activated rhodopsin in nanodiscs. *Proc. Natl Acad. Sci. USA* **114**, E3268–E3275 (2017).
28. Van Eps, N. et al. G_i- and G_s-coupled GPCRs show different modes of G-protein binding. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1721896115 (2018).
29. Oldham, W. M., Van Eps, N., Preininger, A. M., Hubbell, W. L. & Hamm, H. E. Mechanism of the receptor-catalyzed activation of heterotrimeric G proteins. *Nat. Struct. Mol. Biol.* **13**, 772–777 (2006).
30. Oldham, W. M. & Hamm, H. E. Heterotrimeric G protein activation by G-protein-coupled receptors. *Nat. Rev. Mol. Cell Biol.* **9**, 60–71 (2008).
31. Marin, E. P., Krishna, A. G. & Sakmar, T. P. Rapid activation of transducin by mutations distant from the nucleotide-binding site: evidence for a mechanistic model of receptor-catalyzed nucleotide exchange by G proteins. *J. Biol. Chem.* **276**, 27400–27405 (2001).
32. Marin, E. P., Krishna, A. G. & Sakmar, T. P. Disruption of the α5 helix of transducin impairs rhodopsin-catalyzed nucleotide exchange. *Biochemistry* **41**, 6988–6994 (2002).
33. Garcia, P. D., Onrust, R., Bell, S. M., Sakmar, T. P. & Bourne, H. R. Transducin-α C-terminal mutations prevent activation by rhodopsin: a new assay using recombinant proteins expressed in cultured cells. *EMBO J.* **14**, 4460–4469 (1995).
34. Onrust, R. et al. Receptor and $\beta\gamma$ binding sites in the alpha subunit of the retinal G protein transducin. *Science* **275**, 381–384 (1997).
35. Skiba, N. P., Bae, H. & Hamm, H. E. Mapping of effector binding sites of transducin alpha-subunit using $G\alpha_t/G\alpha_{i1}$ chimeras. *J. Biol. Chem.* **271**, 413–424 (1996).
36. Sun, D. et al. Probing $G\alpha_{i1}$ protein activation at single-amino acid resolution. *Nat. Struct. Mol. Biol.* **22**, 686–694 (2015).
37. Itoh, Y., Cai, K. & Khorana, H. G. Mapping of contact sites in complex formation between light-activated rhodopsin and transducin by covalent crosslinking: use of a chemically preactivated reagent. *Proc. Natl Acad. Sci. USA* **98**, 4883–4887 (2001).
38. Subramaniam, S., Gerstein, M., Oesterhelt, D. & Henderson, R. Electron diffraction analysis of structural changes in the photocycle of bacteriorhodopsin. *EMBO J.* **12**, 1–8 (1993).
39. Subramaniam, S. & Henderson, R. Molecular mechanism of vectorial proton translocation by bacteriorhodopsin. *Nature* **406**, 653–657 (2000).
40. Slessareva, J. E. et al. Closely related G-protein-coupled receptors use multiple and distinct domains on G-protein α-subunits for selective coupling. *J. Biol. Chem.* **278**, 50530–50536 (2003).
41. Kling, R. C., Lanig, H., Clark, T. & Gmeiner, P. Active-state models of ternary GPCR complexes: determinants of selective receptor–G-protein coupling. *PloS One* **8**, (2013).
42. DeVree, B. T. et al. Allosteric coupling from G protein to the agonist-binding pocket in GPCRs. *Nature* **535**, 182–186 (2016).
43. Dror, R. O. et al. Signal transduction. Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science* **348**, 1361–1365 (2015).
44. Dickson, B. M., de Waal, P. W., Ramjan, Z. H., Xu, H. E. & Rothbart, S. B. A fast, open source implementation of adaptive biasing potentials uncovers a ligand design strategy for the chromatin regulator BRD4. *J. Chem. Phys.* **145**, 154113 (2016).
45. Huang, W. J. et al. Structural insights into μ-opioid receptor activation. *Nature* **524**, 315–321 (2015).
46. Rose, A. S. et al. Position of transmembrane helix 6 determines receptor G protein coupling specificity. *J. Am. Chem. Soc.* **136**, 11244–11247 (2014).

Acknowledgements Cryo-EM data were collected at the David Van Andel Advanced Cryo-Electron Microscopy Suite in the Van Andel Research Institute. This work was supported in part by the National Institutes of Health grant, DK071662, American Asthma Foundation, Jay and Betty Van Andel Foundation, Ministry of Science and Technology (China) grants 2012ZX09301001 and 2012CB910403, 2013CB910600, XDB08020303, 2013ZX09507001 (to H.E.X.), GM117372 (to A.K.), GM0875119 (to A.A.K.), grant from Pfizer (to A.A.K.), the National Natural Science Foundation 31770796 (to Y.J.), the Canada Excellence Research Chairs program (to O.P.E.), the Canadian Institute for Advanced Research (to O.P.E.), the Anne and Max Tanenbaum Chair in Neuroscience (to O.P.E.), by funds from the Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD (to S.S.), and by federal funds from the Frederick National Laboratory for Cancer Research, National Institutes of Health, under contract HHSN261200800001E. We thank H. Li and W. Lü for help with analysing the cryo-EM data and for advice on refinement, L. Bai and Z. Yuan for advice on 3D reconstruction, V. Falconieri for assistance with figure preparation, the HPC team at VARI for computational support, D. Nadziejka for manuscript editing, and B. Dickson for consultation on molecular dynamics simulation.

Author contributions Y.K. initiated the project, prepared samples, performed data acquisition and structure determination, and prepared the figures and manuscript writing; H.E.X. and K.M. conceived the project and designed the research, and wrote the paper with contributions from all authors; O.K., X.E.Z., A.B. and S.S. performed image processing, structure determination, figure preparation, and manuscript writing; P.W.d.W. performed computational experiments, analysed the structure, prepared figures, and manuscript writing; P.D., S.M., S.E. and A.A.K. designed and performed Fab selection; N.V.E., T.M. and O.P.E. designed and performed DEER experiments; X.G., Y.Y., P.L. and Y.J. performed cell-based assays; G.Z. and X.M. helped with data collection.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0215-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0215-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.A.K., S.S. or H.E.X.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Constructs and expression of human rhodopsin. Human rhodopsin with four mutations, $N2^{Nterm}C$, $E113^{3.28}Q$, $M257^{6.40}Y$ and $N282^{ECL3}C$, was cloned into pFastBac vector. To facilitate expression and purification, an N-terminal 8× His-sfGFP-BRIL epitope and a TEV protease site were inserted after sfGFP. These constructs were expressed in Sf9 insect cells using the Bac-to-Bac Baculovirus system (Invitrogen). The cells were infected with baculovirus at 27 °C for 48 h before collection.

Constructs, expression and purification of G_i heterotrimer. The heterotrimeirc G_i complex was expressed in Sf9 insect cells (Invitrogen). Human $G\alpha_i$ was cloned in pFastbac vector with 6× His-MBP at the N terminus, and the virus was prepared using Bac-to-Bac system (Invitrogen). Rat Gβ1 and N-terminal 6× His-MBP tagged bovine $G\gamma_2$ were cloned into a pFastBac vector. The virus was prepared using the Bac-to-Bac system. The cells were infected with both $G\alpha_i$ and $G\beta\gamma$ virus at 27 °C for 48 h before collection, and the ratio was determined by small-scale titration experiment. The G_i heterotrimer was purified as previously described[26].

DEER spectroscopy. Generation of rhodopsin mutants, expression in HEK293S GnTI$^-$ cells, and spin-labelling with 1-oxyl-2,2,5,5-tetramethyl-Δ^3-pyrroline-3-methyl) Methanethiosulfonate (MTSSL, Toronto Research Chemicals) was as previously described[47]. For DEER measurements, spin-labelled rhodopsin mutants bound to a 1D4-antibody (MA1-722, ThermoFisher)-conjugated column were thoroughly washed and complexed with G_i heterotrimer or with the G_i-Fab_G50 complex in 20 mM HEPES, pH 7.2, 100 mM NaCl, 0.02% *n*-dodecyl-β-D-maltopyranoside (DDM), 0.004% cholesteryl hemisuccinate (CHS) and 5 μM ATR under yellow (>500 nm) light illumination on ice. The entire rhodopsin–G_i or rhodopsin–Gi–Fab_G50 complex was then further washed and eluted from the 1D4 column with 20 mM HEPES, pH 7.2, 100 mM NaCl, 0.1% digitonin, 150 μM 1D4 peptide and 5 μM ATR. The eluate was concentrated, and 20% (v/v) deuterated glycerol was added. The complexes were added to quartz capillaries (1.5 mm inner diameter and 1.8 mm outer diameter) and flash-frozen using a dry ice/ethanol bath. The capillaries were loaded into an EN 5107D2 resonator, and Q-band measurements were performed at 80 K on a Bruker Elexsys 580 spectrometer with a Super Q-FTu Bridge. For the four-pulse DEER experiment used here, a 32-ns π-pump pulse was applied to the low field peak of the nitroxide field-swept spectrum, and the observer π/2 (16 ns) and π (32 ns) pulses were positioned 50 MHz (17.8 G) upfield, which corresponds to the nitroxide centre line. Distance distributions were obtained from the raw dipolar evolution data using the LabVIEW (National Instruments) 'LongDistances' program, developed by C. Altenbach, which can be downloaded from http://www.biochemistry.ucla.edu/Faculty/Hubbell/.

Phage display selections. Avi-tagged heterotrimeric G_i was used for biopanning during selection. The G_i construct has an Avitag on the N-terminal end of the gamma subunit (Gγ). The Avitag was biotinylated in vivo in Sf9 cells by co-expression of the biotin ligase BirA in presence of supplemented D-biotin. Pull-down experiments on Streptavidin magnetic beads showed quantitative biotinylation of the trimeric complex used in selection. Phage display selection was performed according to published protocols[48]. In brief, in the first round, 200 nM of target was immobilized on 250 μl magnetic beads. Then, 100 μl of a phage library[49] containing 10^{12}–10^{13} virions were added to the Streptavidin beads and incubated for 30 min. The resuspended beads containing bound virions were washed extensively and then used to infect freshly grown log phase *Escherichia coli* XL1-Blue cells. Phages were amplified overnight in 2xYT media (Fisher Scientific) with 50 μg ml^{-1} ampicillin and 10^9 p.f.u. ml^{-1} of M13-KO7 helper phage. To increase the stringency of selection, three additional rounds of sorting were performed with decreasing the target concentration in each round (second round: 50 nM, third round: 10 nM and fourth round: 10 and 5 nM) using the amplified pool of virions of the preceding round as the input. Sorting from the second to fourth rounds was done on a Kingfisher instrument. From the second to fourth rounds, the targets were premixed with the amplified phage pool and then Streptavidin beads were added to the mixture. From the second round onwards, the bound phages were eluted using 0.1 M glycine, pH 2.7. This technique often risks the elution of non-specific and Streptavidin binders, which tend to overpopulate the amplified phage pool thereby reducing the chance to obtain the desired specific clones. To eliminate them, the precipitated virions from the second round onwards were negatively selected against 100 μl of Streptavidin beads. The pre-cleared phage pool was then used as an input for the selection.

Single-point phage ELISA. All ELISA experiments were performed in 96-well plates coated with 50 μl of 2 μg ml^{-1} neutravidin in Na_2CO_3 buffer, pH 9.6 and subsequently blocked by 0.5% BSA in PBS. A single-point phage ELISA was used to rapidly screen the binding of the obtained Fab fragments in phage format. Colonies of *E. coli* XL1-Blue harbouring phagemids were inoculated directly into 500 μl of 2xYT broth supplemented with 100 μg ml^{-1} ampicillin and M13-KO7 helper phage. The cultures were grown at 37 °C for 16–20 h at 280 r.p.m. in a 96-deep-well block plate. Culture supernatants containing Fab phage were diluted tenfold in PBST buffer. After 15 min of incubation, the mixtures were transferred to ELISA plates that were incubated with 50 nM biotinylated trimeric G_{i1} in experimental wells and with buffer in control wells for 15 min. The ELISA plates were incubated with the phage for another 15 min and then washed with PBST. The washed ELISA plates were incubated with horseradish perixoidase (HRP)-conjugated anti-M13 mouse monoclonal antibody (ab50370, Abcam,1:5,000 dilution in PBST) for 30 min. The plates were again washed, developed with TMB substrate and then quenched with 1.0 M HCl, and the absorbance at 450 nm was determined. The background binding of the phage was monitored by the absorbance from the control wells.

Sequencing, cloning, overexpression and purification of Fab fragments. From phage ELISA, clones (selected based on a high ratio of ELISA signal of target binding to background) were sequenced at the DNA Sequencing Facility at the University of Chicago. Unique clones were sub-cloned in pRH2.2, an IPTG inducible vector for expression of Fabs in *E. coli*. *E. coli* BL21 (Gold) cells were transformed with sequence-verified clones of Fab fragments in pRH2.2. Fab fragments were grown in 2xYT media with 100 μg ml^{-1} ampicillin at 37 °C for 2–2.5 h, during which $A_{600\,nm}$ reached 0.6–0.8, induced with 1 mM IPTG and grown for a further 4.5 h at 37 °C. Harvested cells were kept frozen at −80 °C until use. Frozen pellets were re-suspended in PBS supplemented with 1 mM PMSF, and 1 μg ml^{-1} DNase I. The suspension was lysed by ultrasonication. The cell lysate was incubated at 65 °C for 30 min to eliminate of any undesired proteolysed fragments of the Fab produced during overexpression. Heat-treated lysate was then cleared by centrifugation, filtered through 0.22 μm filter and loaded onto a HiTrap MabSelect SuRe 5-mL column pre-equilibrated with lysis buffer (20 mM phosphate buffer, pH 7.5, 500 mM NaCl). The column was washed with 10 column volumes of lysis buffer followed by elution of Fab fragments with elution buffer (0.1 M acetic acid). Fractions containing protein were directly loaded onto a Resource S 1-mL column pre-equilibrated with buffer A (50 mM sodium acetate, pH 5.0) followed by washing with 10 column volumes wash with buffer A. Fab fragments were eluted with a linear gradient 0–50% of buffer B (50 mM sodium acetate, pH 5.0, 2.0 M NaCl). Affinity and ion-exchange chromatography were performed using an automated program on ÄKTA explorer system. Purified Fab fragments were dialysed overnight against 20 mM HEPES, pH 7.4, 150 mM NaCl. The quality of purified Fab fragments was analysed by SDS–PAGE.

Rhodopsin–G_i–Fab complex formation and purification. Sf9 cell pellets infected with virus containing rhodopsin were lysed in 20 mM HEPES, pH 7.2, 10 mM NaCl, and 10 mM $MgCl_2$. The supernatant was centrifuged at 160,000*g* for 30 min to collect the membranes. The membranes were washed by homogenization in 20 mM HEPES, pH 7.2, 1 M NaCl and 10 mM $MgCl_2$, and then was collected by centrifugation at 160,000*g* for 30 min.

The rhodopsin–G_i complex was formed in membranes as described previously[15]. The washed membranes were re-suspended in 20 mM HEPES, pH 7.2, 100 mM NaCl, 10% glycerol, and 5 μM ATR. For 1 l of rhodopsin cell pellets, 6 mg of G_i 10 mg Fab_G50 and 1 U of apyrase were added. The sample was incubated overnight at 4 °C. The membranes were then solubilized in 20 mM HEPES, pH 7.2, 100 mM NaCl, 10% glycerol, 0.5% DDM (Anatrace), 0.1% cholesteryl hemisuccinate (CHS), and 5 μM ATR for 2 h at 4 °C. The supernatant was isolated by centrifugation at 160,000*g* for 1 h, and then was incubated with TALON IMAC resin (Clontech) for 3 h at 4 °C. After binding, the resin was washed with 10 column volumes of 20 mM HEPES, pH 7.2, 100 mM NaCl, 0.02% DDM, 0.004 CHS, 50 mM imidazole and 5 μM ATR. The buffer was exchanged to 20 mM HEPES, pH 7.2, 100 mM NaCl, 0.1% digitonin and 5 μM ATR. The protein was then treated overnight with His-tagged TEV protease (made in-house) on column. The complex sample was eluted with 20 mM HEPES, pH 7.2, 100 mM NaCl, 0.1% digitonin, 50 mM imidazole and 5 μM ATR. The rhodopsin–G_i–Fab complex sample was concentrated and loaded onto Superdex S200 10/300 GL column with running buffer 20 mM HEPES, pH 7.2, 100 mM NaCl, 0.1% digitonin and 5 μM ATR; the fractions for the monomeric complex were collected and concentrated individually for electron microscopy experiments.

Negative-stain analysis of rhodopsin–G_i–Fab complex. Protein samples were applied to a freshly glow-discharged carbon coated copper grid and allowed to adhere for 10 s before being reduced to a thin film by blotting. Immediately after blotting, 3 μl of a 1% solution of uranyl formate was applied to the grid and blotted off directly. This was repeated three times. Data were acquired using a Tecnai Spirit transmission electron microscope operating at 120 kV. Images were processed using Relion 2.1[50].

Cryo-EM data acquisition. A droplet (2.75 μl) of purified rhodopsin–G_i–Fab complex at a concentration of about 9 mg ml^{-1} was applied to a glow-discharged holey carbon grid (Quantifoil R1.2/1.3, Au 300 mesh), and subsequently vitrified using a Vitrobot Mark IV (FEI Company). Cryo-EM data were collected on a Titan Krios

microscope using a K2 camera positioned post a GIF quantum energy filter, with a slit width of 20 eV. Micrographs were recorded in super-resolution mode at a magnified physical pixel size of 1.074 Å, with defocus values ranging from −1.3 to −3.0 μm. The total exposure time was set to 6 s with intermediate frames recorded every 0.2 s, resulting in an accumulated dose of about 60 electrons per Å^2 and a total of 30 frames per movie stack.

Image processing and structural refinement. In the initial phase of processing, unbinned image stacks from 19,368 K2 movies were corrected for drift and for beam-induced motion by alignment using cross-correlation as implemented in Unblur[51]. Dose-corrected integrated frames were used for subsequent image processing. Particles were automatically picked using RELION 2.1[50] using 6 projection image references, generated by applying 2D classification on 1,000 manually picked particles. CTFFIND4 was used for CTF estimation[52]. Next, a set of 1.65 million particles were extracted from the 14,464 integrated frames displaying Thon rings extending beyond 4.5 Å. The extracted particle images were normalized, and subjected to 50 rounds of both iterative 2D classification (regularization parameter $T=2$) and 3D classification ($T=4$). Uninterpretable, sparsely populated, or poorly defined classes were discarded at both stages leaving behind approximately 227 k particles for further 3D processing (Extended Data Fig. 3). A coarse initial model was generated using the 3D initial model generation module in RELION and the model was refined until convergence was achieved. Density maps were corrected with a *B*-factor of −217 and 'gold-standard' Fourier shell correlation (FSC) resolution plots were calculated with a soft shape mask applied to independent, unfiltered half-maps resulting from the processing.

The crystal structure of human rhodopsin (PDB code 4ZWJ) and G protein complex (PDB code 1GG2) were used as initial models for model rebuilding and refinement against the electron microscopy map. All models were docked into the electron microscopy density map using Chimera[53], followed by iterative manual adjustment in COOT[54], fragment-based refinement with Rosetta[55], and real space refinement using Phenix programs[56]. The model statistics was validated using MolProbity[57]. Structural figures were prepared in Chimera and PyMOL (https://pymol.org/2/). The final refinement statistics are provided in Supplementary Table 3. The extent of any model overfitting during refinement was measured by refining the final model against one of the half-maps and by comparing the resulting map versus model FSC curves with the two half-maps and the full model.

In-cell disulfide bond cross-linking. The open-reading frames of mini-Gi with an N-terminal Flag tag and full-length rhodopsin with a C-terminal haemagglutinin (HA) tag were cloned into pcDNA6. Cysteine mutations (E28C for miniG$_i$ and N145C or F146C for rhodopsin) were systematically introduced in these two DNA vectors. AD293 cells were split 1 day before transfection at 50,000 cells per well in a 24-well plate. Cells were grown for 1 day, then transfected with 100 ng rhodopsin constructs (pcDNA6-rho-3HA) plus 100 ng G$_i$ plasmid (pcDNA6-3xFlag-miniGi), 100 ng pcDNA6-Gβ, and 100 ng pcDNA6-Gγ by Lipofectamine 2000 (DNA:Lipofectamine 2000 ratio of 1:2) in each well. Cells were grown for 2 days after transfection, and were then treated at room temperature with H_2O_2, which was freshly diluted in the cell culture medium to a final concentration of 1 mM. After 5 min of treatment with H_2O_2, the medium was aspirated and 100 μl of CelLytic M (Sigma C2978) were added to each well and the plate was shaken for 10 min at room temperature. Cell lysates were transferred to a 1.5 ml tubes, spun at 18,000*g* at 4 °C for 5 min. The supernatants (10 μl) were mixed with an equal volume of 2× SDS loading buffer (without reducing agents) for 5 min at room temperature, and loaded onto a protein gel for western blot analysis. HRP-conjugated anti-Flag (A8592, Sigma) and anti-HA (H3663, Sigma) antibodies were used to probe for free and cross-linked miniGi and rhodopsin proteins.

Gα$_i$-mediated ERK activation assay and cAMP accumulation assay. AD293 cells were plated at a density of 5×10^4 per well in 24-well plates 1 day before transfection. Cells were then transiently transfected using Lipofectamine 2000 reagent (Life Technologies) with 50 ng cDNA encoding GPCR, 200 ng luciferase reporter construct containing an ERK response element (SRE) or cAMP response element (CRE), and 10 ng TK-*Renilla* at a Lipofectamine 2000 reagent:DNA ratio of 2:1. ERK activation or cAMP accumulation was detected using the Dual-luciferase reporter assay system from Promega according to the manufacturer's instructions using an EnVision plate reader (Perkin Elmer). *Renilla* luciferase was used for normalization. All experiments were performed in triplicate, with each well transfected independently.

Molecular dynamics simulation setup and equilibration. Molecular dynamics simulations of two G$_i$ couplers, rhodopsin (PDB code 3PQR) and mOR1 (PDB code 5C1M), and two G$_s$ couplers, β_2AR (PDB code 4LDE) and A_{2A}R (PDB code 5G53), were initiated from structures in the active (S4) conformation as described previously[58]. All receptors selected met the following inclusion criteria: a fully active, agonist-bound receptor in complex with a peptide, nanobody, or engineered G protein, and free of any ICL3 fusion that may influence the conformational state of the receptor.

Each receptor was prepared for simulations as follows. Crystallization partners and heteroatoms, with exception of agonists bound within the orthosteric binding site and palmitoylated cysteine residues on helix 8, were removed. Thermostabilizing and non-native mutations were reverted back to their wild-type residue and missing residues not within ICL3 were modelled and subjected to 5,000 rounds of 'very_slow' loop refinement assed by DOPE scoring using Modeller 9.17[59]. Palmitoyl groups on helix 8 were then added and protonation states for active receptors were assigned based on previous publications[45,60,61]. The resulting complexes were capped with neutral acetyl and methylamine groups and embedded into a pre-equilbirated palmitoyl-oleoyl-phosphatidylcholine (POPC) lipid bilayer, solvated in a box of TIP3P waters allowing for 14 Å of padding on all sides with 150 mM NaCl, and neutrailized by removing appropriate ions or counter ions using the Desmond system builder within Maestro (Schrödinger Release 2018-1: Maestro, Schrödinger, LLC, New York, 2018). Full details for each system can be found in Supplementary Table 1.

All-atom atmospheric simulations were performed using GROMACS5.0.6 with the CHARMM36m force field and periodic boundary conditions[62]. Ligand parameters were generated by SwissParam[63]. Before production simulations, 50,000 steps of energy minimization were performed, followed by equilibration in the canonical (NVT) and isothermal–isobaric (NPT) ensembles for 10 and 50 ns, respectively, with positional restraints (1,000 kJ mol^{-1} nm^{-2}) placed on backbone atoms. Temperature was maintained at 310 K using the v-rescale method with a coupling time of 0.1 ps and pressure was maintained at 1 bar using the Berendsen barostat with a coupling time (t_p) of 1.0 ps and compressibility of 4.5×10^{-5} bar^{-1} with a 2 fs timestep[64].

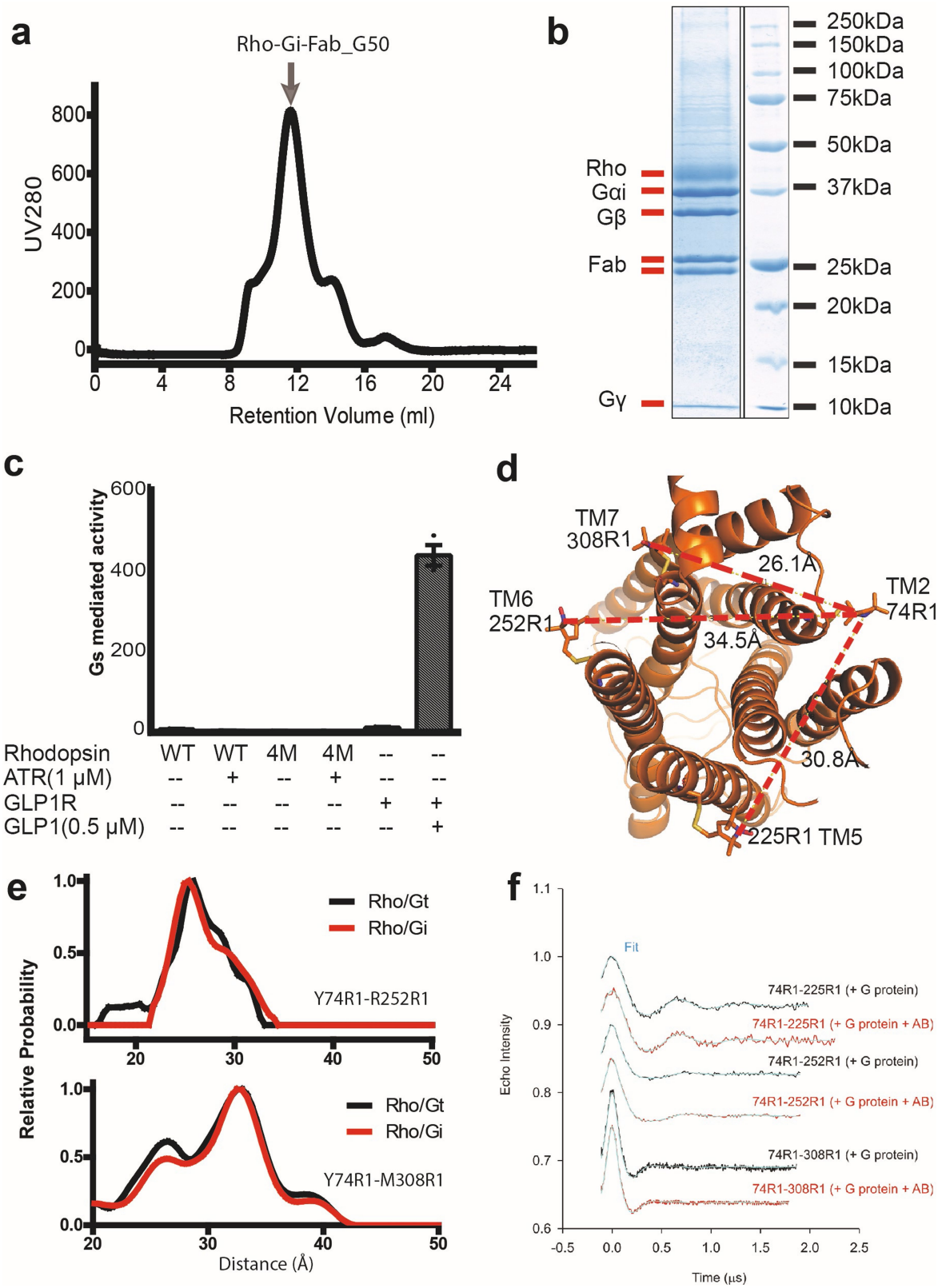
Mollified adaptive biasing potential simulations. To calculate the free energy landscape of TM6 movement relative to the transmembrane bundle of the receptor, we used mABP with overfill protection using fABMACS[65,66]. A centre-of-geometry (COG) distance collective variable was introduced into fABMACS and two collective variables were defined: COG distance between select TM6 (6.27–6.42) and TM1/2 (1.50–1.59 and 2.39–2.48) C$_\alpha$ atoms, and COG distance between select TM6 and TM3/4 (3.45–3.55 and 4.39–4.47) C$_\alpha$ atoms (Extended Data Fig. 7). The collective variable space was discretized into a 480 × 480 grid with ranges of 0 to 50 Å for a bin width of 0.10 Å. Biasing parameters were $b=0.8$, $c=0.01/\delta t$, $\alpha=10$, and a maximum fill level of 30 kJ mol^{-1}. Simulation parameters for mABP simulations were the same as those used during NPT equilibration except the Parrinello–Rahman barostat with a 5.0 ps coupling time was used. To prevent unwanted transitions away from the active conformation of the receptors, the NPXXY tyrosine was held in an active conformation and TM7 was restrained using Urey–Bradley harmonic potentials between residues 7.42–7.53, in which angles and distances were derived from the equilibrated structure and residues *i*, *j*, *k* of the potential were *i*, *i* + 2, *i* + 4. Simulations without TM7 restraints saw quick transitions towards receptor inactivation, consistent with prior long time scale simulations[67] (data not shown). For simulations in which ICL3 was truncated or not resolved, backbone distance restraints were placed on the last four residues to prevent helical unwinding. Independent mABP production simulations for each receptor were run in duplicate for approximately 12 μs in total. Throughout the simulations, the instantaneous boost was recorded for each recorded frame allowing for the generation of weighted histograms. Simulation analysis was performed using MDTraj 1.7.2 and VMD 1.9.2[1,7]. Plots were generated using the R statistical package.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

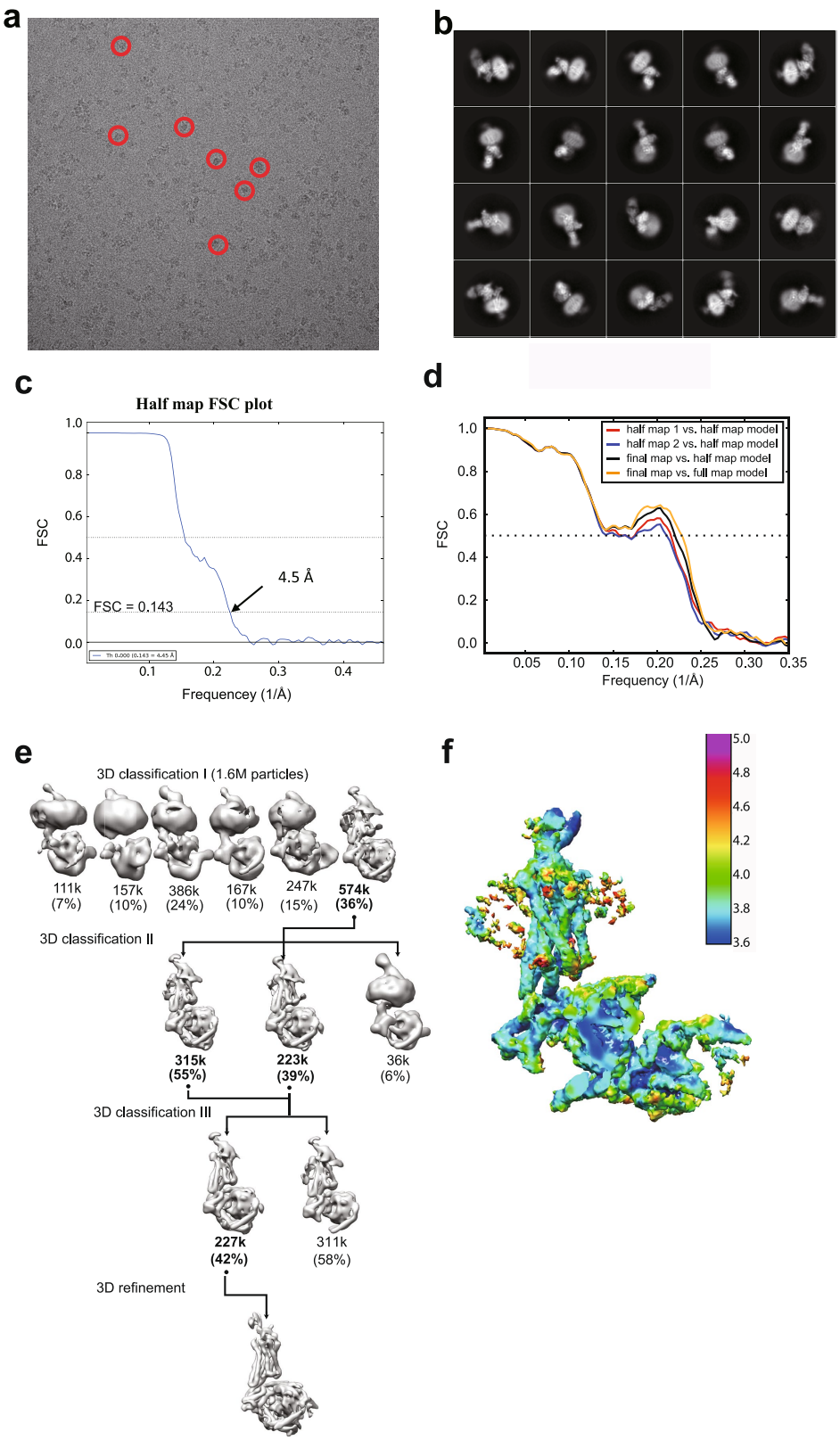
Data availability. All data and source code are available upon request.

47. Caro, L. N. et al. Rapid and facile recombinant expression of bovine rhodopsin in HEK293S GnTI$^-$ cells using a *PiggyBac* inducible system. *Methods Enzymol.* **556**, 307–330 (2015).
48. Hornsby, M. et al. A high through-put platform for recombinant antibodies to folded proteins. *Mol. Cell Proteomics* **14**, 2833–2847 (2015).
49. Paduch, M. & Kossiakoff, A. A. Generating conformation and complex-specific synthetic antibodies. *Methods Mol. Biol.* **1575**, 93–119 (2017).
50. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
51. Brilot, A. F. et al. Beam-induced motion of vitrified specimen on holey carbon film. *J. Struct. Biol.* **177**, 630–637 (2012).
52. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
53. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
54. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
55. Wang, R. Y. et al. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *eLife* **5**, e17219 (2016).
56. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).

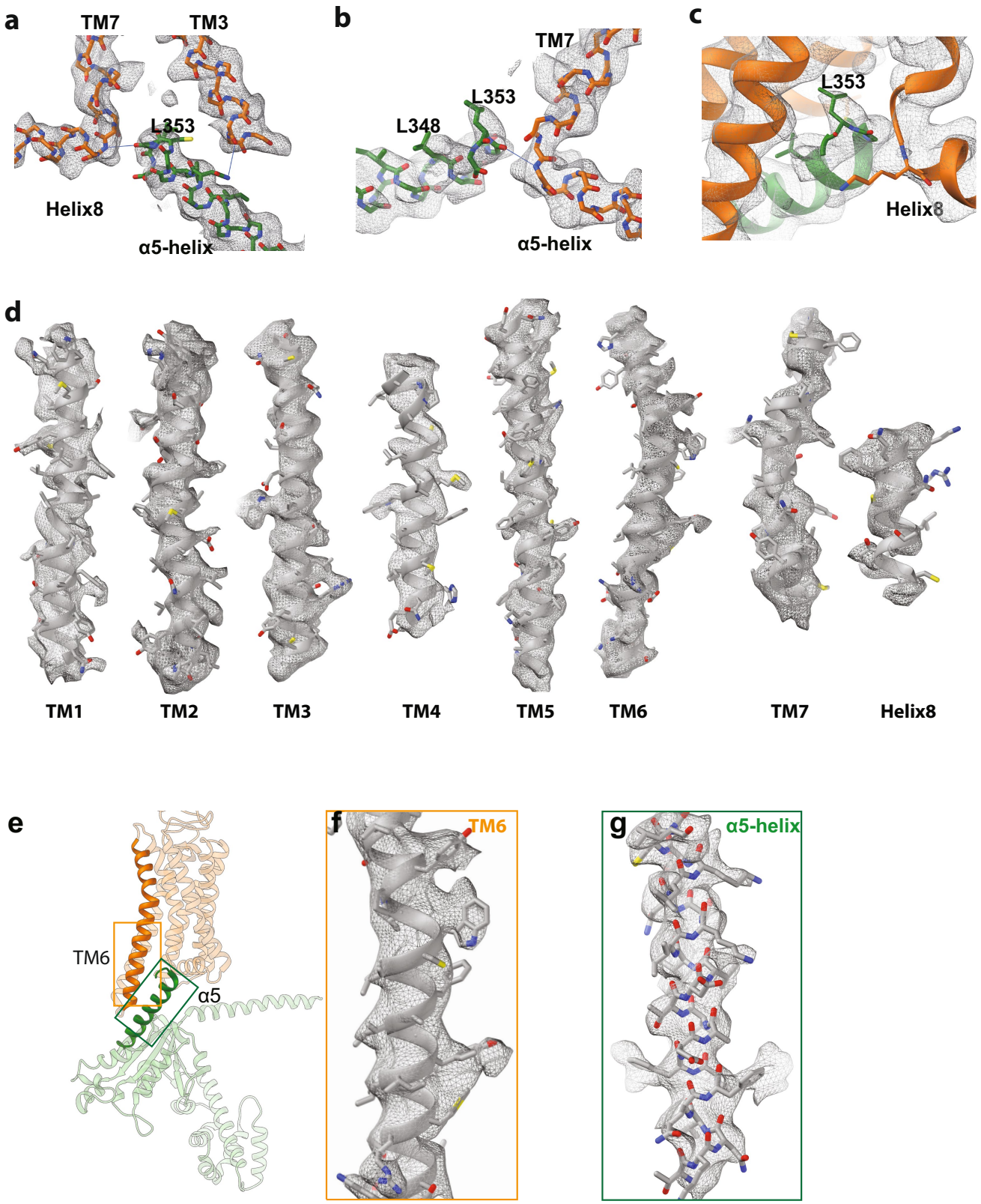
57. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
58. Manglik, A. et al. Structural insights into the dynamic process of β_2-adrenergic receptor signaling. *Cell* **161**, 1101–1111 (2015).
59. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*. **234**, 779–815 (1993).
60. Mahalingam, M., Martínez-Mayorga, K., Brown, M. F. & Vogel, R. Two protonation switches control rhodopsin activation in membranes. *Proc. Natl Acad. Sci. USA* **105**, 17795–17800 (2008).
61. Ranganathan, A., Dror, R. O. & Carlsson, J. Insights into the role of $Asp79^{2.50}$ in β_2 adrenergic receptor activation from molecular dynamics simulations. *Biochemistry* **53**, 7283–7296 (2014).
62. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
63. Zoete, V., Cuendet, M. A., Grosdidier, A. & Michielin, O. SwissParam: a fast force field generation tool for small organic molecules. *J. Comput. Chem*. **32**, 2359–2368 (2011).
64. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys*. **126**, 014101 (2007).
65. Dickson, B. M., de Waal, P. W., Ramjan, Z. H., Xu, H. E. & Rothbart, S. B. A fast, open source implementation of adaptive biasing potentials uncovers a ligand design strategy for the chromatin regulator BRD4. *J. Chem. Phys*. **145**, 154113 (2016).
66. Dickson, B. M. Overfill protection and hyperdynamics in adaptively biased simulations. *J. Chem. Theory Comput*. **13**, 5925–5932 (2017).
67. Kohlhoff, K. J. et al. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem*. **6**, 15–21 (2014).
68. Alexander, S. P. H. et al. The Concise Guide To Pharmacology 2017/18: G protein-coupled receptors. *Br. J. Pharmacol*. **174** (Suppl. 1), S17–S129 (2017).



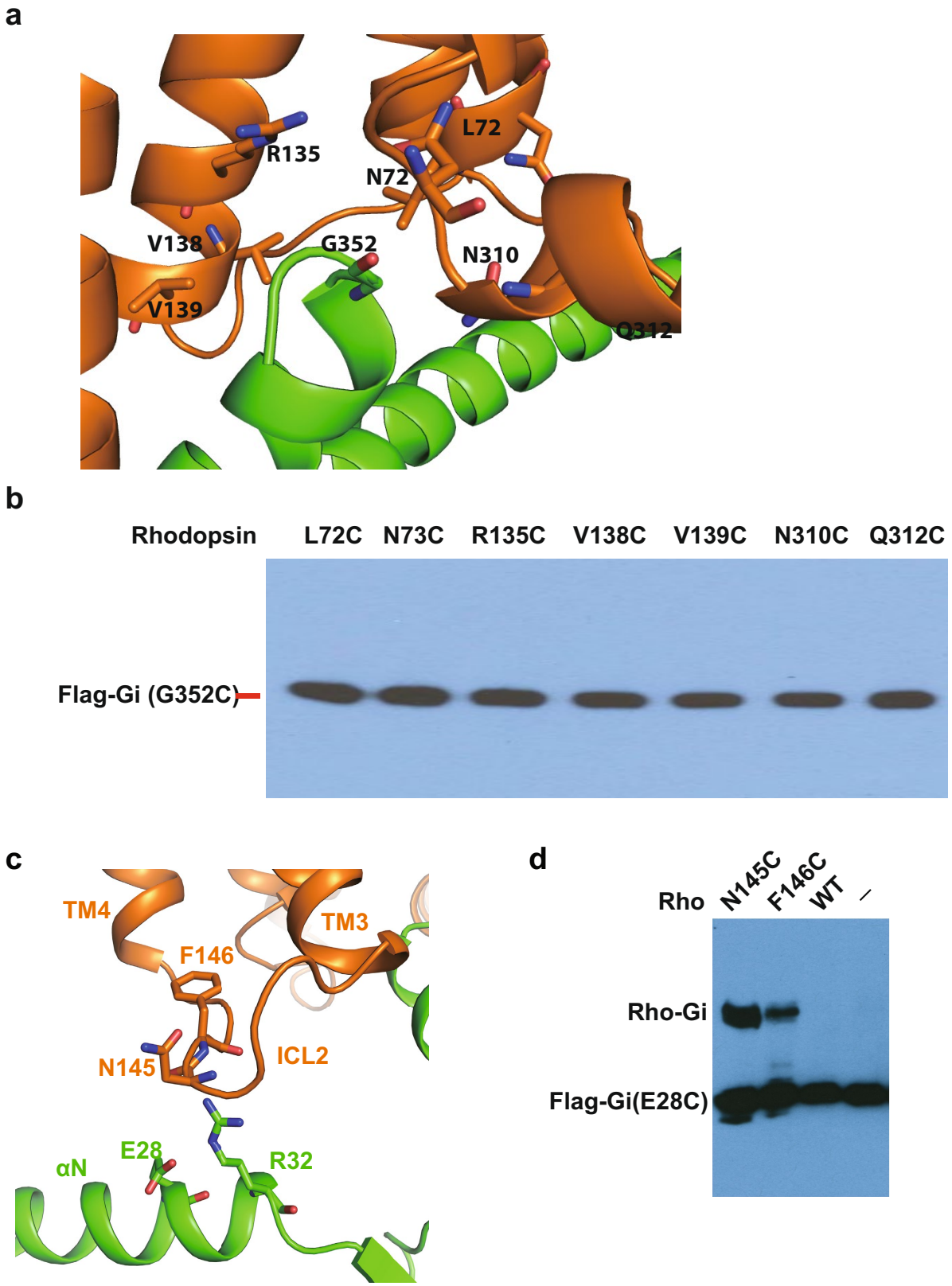
Extended Data Fig. 1 | Purification, characterization and cryo-EM images of the Rho–G_i–Fab complex. a, Representative elution profile of the purified Rho–G_i–Fab_G50 complex on Superdex 200 10/300 gel filtration. **b**, SDS–PAGE analysis of the complex after gel filtration. **c**, The inability of rhodopsin to stimulate the G_s-mediated signalling as assayed by the cAMP-driven luciferase reporter assays. The glucagon-like peptide 1 receptor (GLP-1R) shows stronger G_s-meditated signalling with the agonist GLP-1 ($n = 3$ independent experiments). Data are mean ± s.d. **d**, An overall view of rhodopsin showing the three intramolecular distances between two nitroxide N–O bonds based on the models of the R1 nitroxide pairs Y74R1-Q225R1, Y74R1-R252R1 and Y74R1-M308R1, respectively ($Y74^{2.41}$, $Q225^{5.60}$, $R252^{6.35}$, $M308^{7.55}$; superscripts denote Ballesteros–Weinstein numbering). R1 side-chain modelling details have been described previously[27]. **e**, Similar DEER distance distributions of TM6 and TM7 to TM2 of rhodopsin bound to G_i and G_t. **f**, Time domain data of DEER measurements.



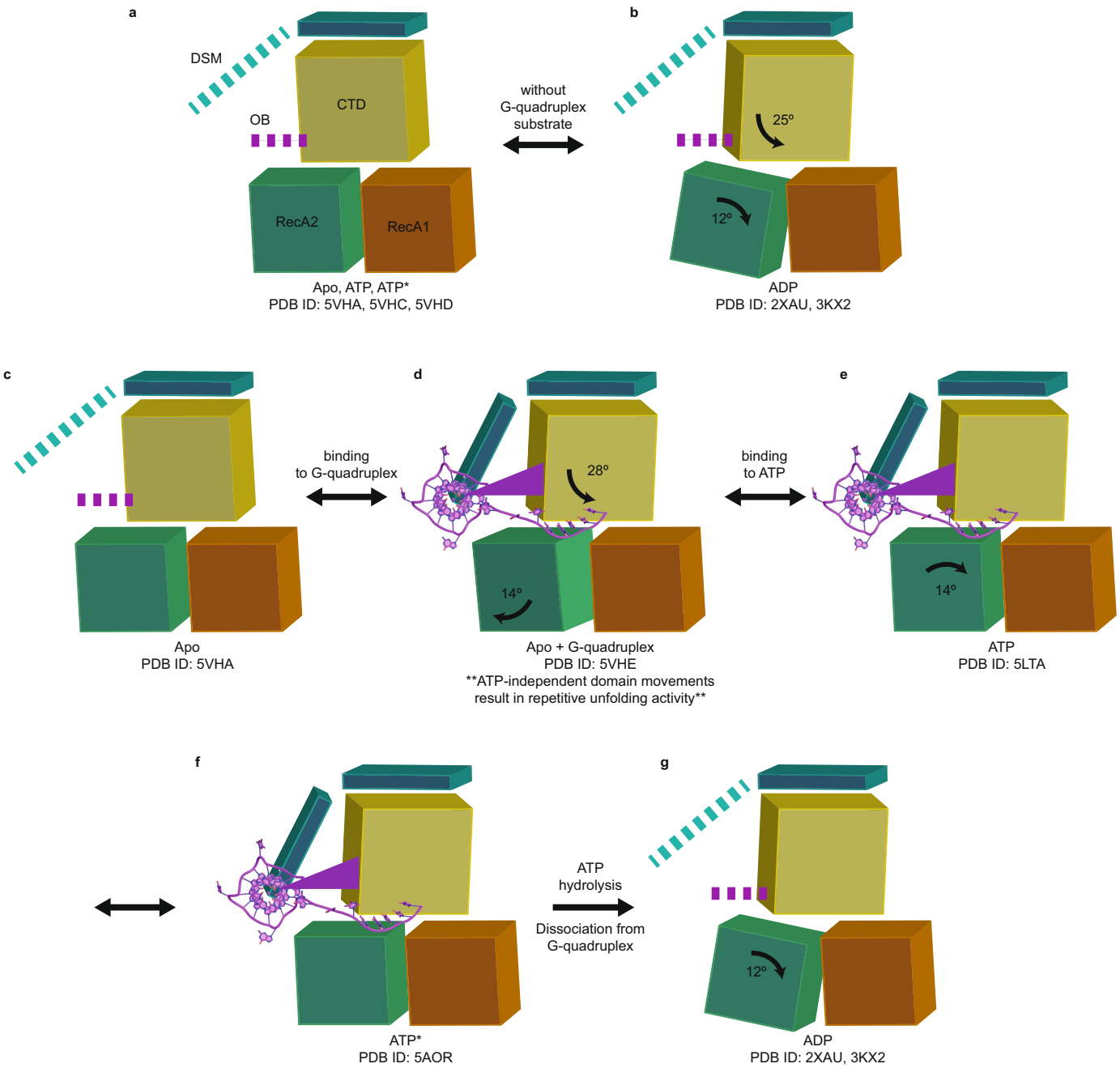
Extended Data Fig. 2 | Cryo-EM images and single-particle analysis of the Rho–G_i–Fab complex. a, Representative cryo-EM micrograph of Rho–G_i–Fab complex. Examples of particle projections are circled. **b**, Reference-free two-dimensional class averages of the complex in digitonin micelles. **c**, Half-map Fourier shell correlation (FSC) plots as produced by RELION with the mask used shown as an inset. **d**, FSC curve of model versus the full map, as well as FSC curves obtained for a model refined against a half-map and compared to the two half-maps as well as the full model. The r.m.s.d. between the model refined against half-map and compared to the full map, and the one refined against the full map is 0.984 Å, and their corresponding FSCs against the final map show a resolution difference at the 0.5-cutoff of approximately 0.1 Å. **e**, Particle classification and refinement. **f**, Local resolution map of the rhodopsin–G_i complex.



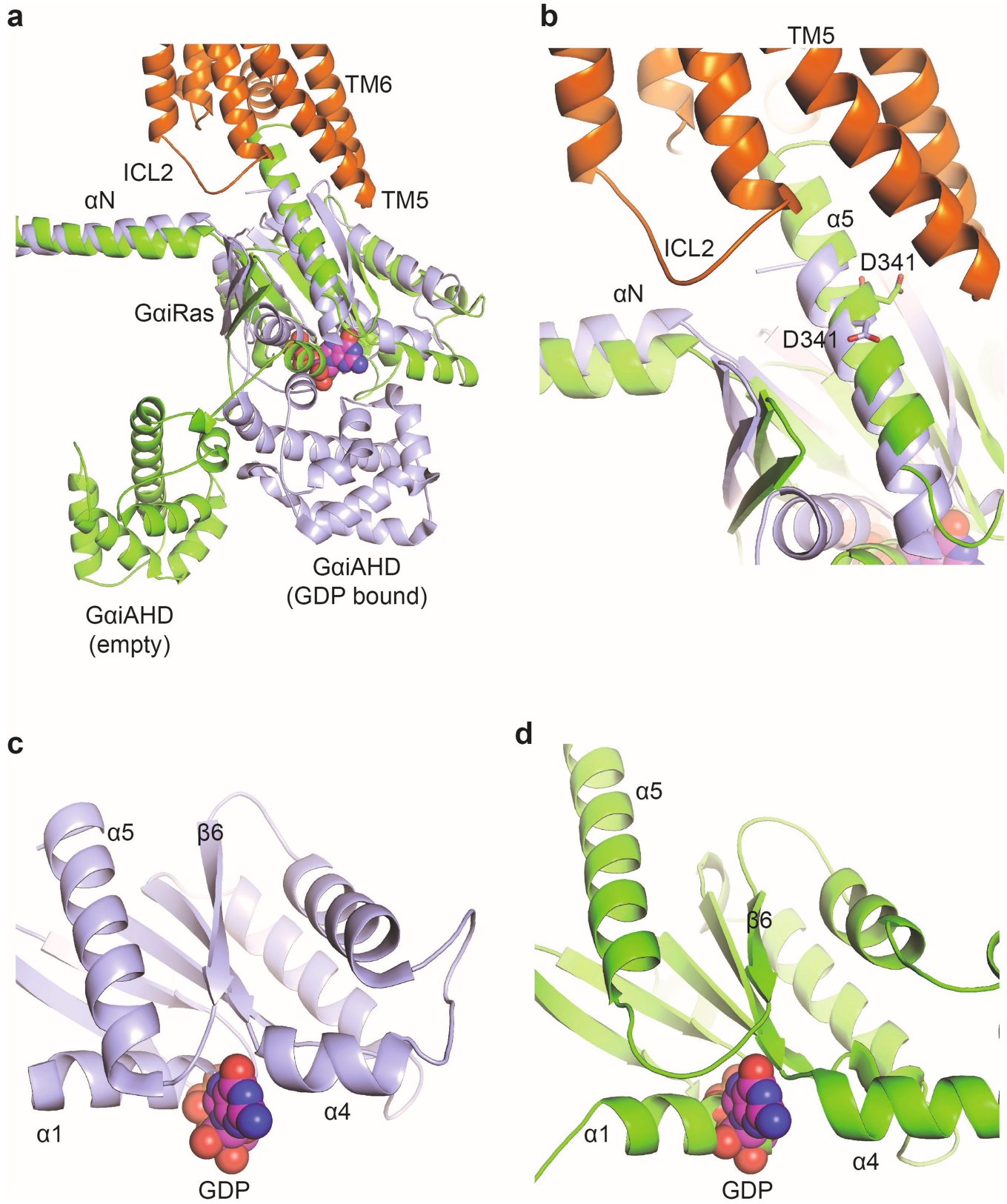
Extended Data Fig. 3 | Electron microscopy density map of rhodopsin–G_i complex. a–c, Three views of the electron microscopy density map of the rhodopsin–$G\alpha_i$ interface. **d**, Electron microscopy density map of all rhodopsin transmembrane helices and helix 8. **e–g**, An overall view of the rhodoposin–$G\alpha_i$ interface (**e**), and electron microscopy density map for the TM6 of rhodopsin (**f**) and the α5-helix of $G\alpha_i$ (**g**).



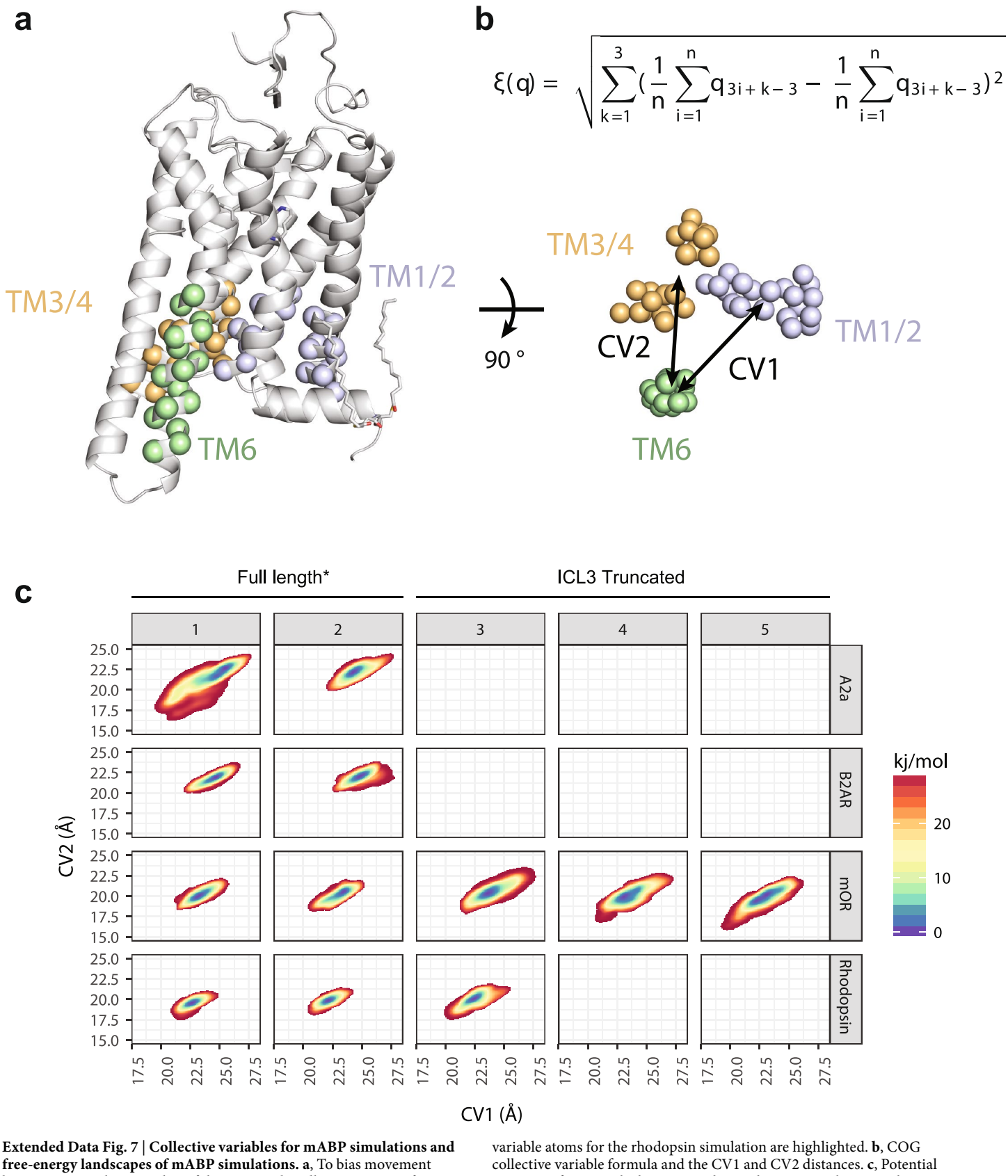
Extended Data Fig. 4 | The rhodopsin–G_i interface and disulfide crosslinking of rhodopsin with $G\alpha_i$. a, The rhodopsin–G_i interface surrounding the G352 residue of $G\alpha_i$ α5-helix. Not all side chains shown are visible in the map but shown here for illustrating their C_α positions to facilitate understanding of data in panel **b**. **b**, Lack of disulfide crosslinking of G352C of G_i with surrounding residues from rhodopsin (compare with **d**; $n = 3$ independent experiments). **c**, Interactions at the interface between ICL2 of rhodopsin and αN helix of $G\alpha_i$. The side chains are not visible in the map but shown here for illustrating their C_α positions. **d**, Demonstration that E28C of $G\alpha_i$ can be disulfide cross-linked to rhodopsin residues $N145C^{ICL2}$ and $F146C^{ICL2}$ ($n = 3$ independent experiments).



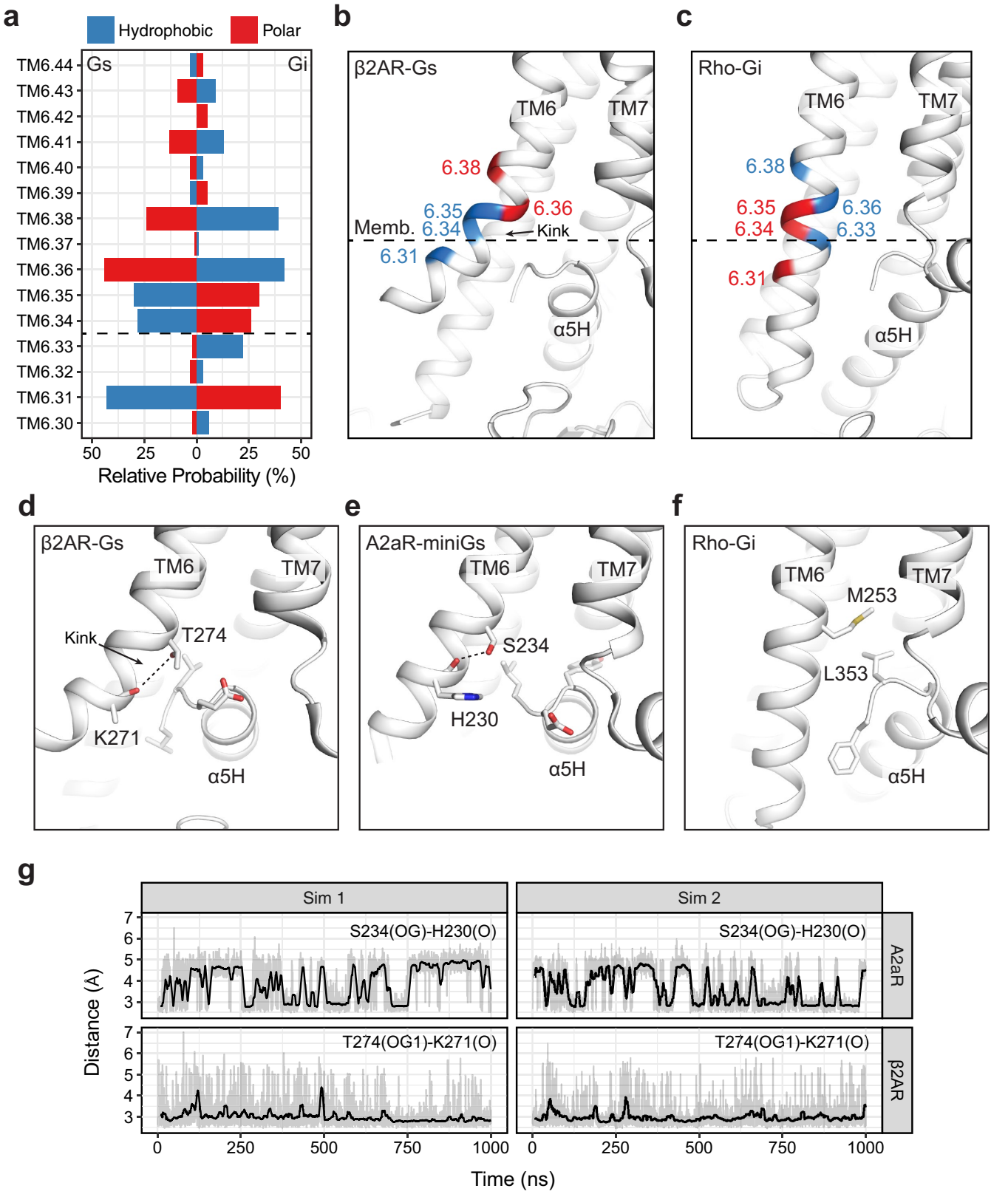
Extended Data Fig. 5 | Structural comparison of G_i-bound rhodopsin, G_s-bound GLP-1R, and G_s-bound CTR, and the role of α4-helix of Gα in receptor selectivity. **a**, **b**, Side and cytoplasmic views of G_i-bound rhodopsin (orange) overlaid with G_s-bound GLP-1R (PDB code 5VAI, light blue, black arrows indicate differences in helix positions). **c**, **d**, Side and cytoplasmic views of G_i-bound rhodopsin (orange) overlaid with G_s-bound CTR (PDB code 5UZ7, grey). **e**, **f**, Side-by-side comparison of the rhodopsin–G_i complex (**e**) with the β_2AR–G_s complex (**f**). **g**. An overlay of the rhodopsin–G_i complex with the β_2AR–G_s complex reveals possible collision of TM5 of β_2AR with α4-helix of $G\alpha_i$.



Extended Data Fig. 6 | The mechanism of rhodopsin-mediated G_i activation. a, **b**, Superposition of the rhodopsin–G_i complex with the inactive GDP-bound G_i (PDB code 1GG2) reveals separation of the AHD from the Ras domain of $G\alpha_i$ (**a**) and conformational changes in the α5-helix (**b**). **c**, **d**, Side-by-side comparison of the GDP-binding site of the $G\alpha_i$ Ras domain in the inactive GDP-bound $G\alpha_i$ (**c**) and nucleotide-free state $G\alpha_i$ with GDP added for comparison (**d**).



Extended Data Fig. 7 | Collective variables for mABP simulations and free-energy landscapes of mABP simulations. a, To bias movement between TM6 relative to that of the receptor bundle, two centre-of-geometry (COG) distance collective variables (CVs) were implemented into fABMACS[66]. CV1 and CV2 are COG distances between selected atoms of TM6 to TM1/2 and TM6 to TM3/4 respectively. Collective variable atoms for the rhodopsin simulation are highlighted. **b**, COG collective variable formula and the CV1 and CV2 distances. **c**, Potential energy surface reveals that CV1 and CV2 distances are larger in the G_s-coupled receptors ($A_{2A}R$ and β_2AR) than those in the G_i-coupled receptors (mOR1 and rhodopsin).



Extended Data Fig. 8 | Enrichment profiles for G_i and G_s coupling receptors. a–c, Relative probability of hydrophobic and polar residues for G_i ($n = 76$) and G_s ($n = 25$) coupling receptors. Residues with relative enrichments over 20% were mapped onto the structures of G_s-bound β_2AR (**b**) and G_i-bound rhodopsin (**c**). GPCR principal coupling was previously defined[68]. **d–f**, Interaction network of TM6.36 of β_2AR, A_{2A}R and rhodopsin with the G protein α5-helix. **g**, Hydrogen bonding between TM3.36 and the backbone of TM6.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	#1 Rho-Gi (EMDB-7517) (PDB 6CMO)
Data collection and processing	
Magnification	45956
Voltage (kV)	300
Electron exposure (e–/Å^2)	60
Defocus range (μm)	-1.5 to -2.5
Pixel size (Å)	1.074
Symmetry imposed	C1
Initial particle images (no.)	1656874
Final particle images (no.)	227386
Map resolution (Å)	4.5
FSC threshold	0.5
Map resolution range (Å)	3.6-5
Refinement	
Initial model used (PDB code)	
Model resolution (Å)	4.5
FSC threshold	0.143
Model resolution range (Å)	80-4.5
Map sharpening *B* factor (Å^2)	-217
Model composition	
Non-hydrogen atoms	11835
Protein residues	1518
Ligands	2
B factors (Å^2)	
Protein	129.9
Ligand	297.5
R.m.s. deviations	
Bond lengths (Å)	0.004
Bond angles (°)	1.001
Validation	
MolProbity score	1.23
Clashscore	4.5
Poor rotamers (%)	0.2
Ramachandran plot	
Favored (%)	98.67
Allowed (%)	1.33
Disallowed (%)	0

Extended Data Table 2 | GPCR simulation systems used in the current study

Receptor (Coupling)	rhodopsin (Gi/t)	μ-opioid receptor 1 (Gi)	β2 adrenergic receptor (Gs)	adenosine A2A receptor (Gs)
PDB (resolution)	3PQR (2.85 Å)	5C1M (2.10 Å)	4LDE (2.79 Å)	5G53 (3.40 Å)
Crystallization Partner	Gt peptide	nanobody	nanobody	mini Gs
Ligand	ATR (agonist)	BU72 (agonist)	BI167107 (Agonist)	NEC (agonist)
System Size	77.2 Å x 77.2 Å x 97.0 Å; 10,610 waters; 128 POPC; 54,353 atoms total	79.1 Å x 79.1 Å x 93.3 Å; 10,347 waters; 135 POPC; 54,171 atoms total	81.1 Å x 81.1 Å x 89.2 Å; 10,198 waters; 143 POPC; 54,557 atoms total	90.2 Å x 90.2 Å x 89.7 Å; 12,720 waters; 191 POPC; 68,568 atoms total
Protonated Residues	$D83^{2.50}$, $E113^{3.28}$, $E122^{3.37}$, $E134^{3.49}$	$D116^{2.50}$	$D79^{2.50}$, $E122^{3.41}$, $D130^{3.49}$	$D52^{2.50}$,$D101^{3.49}$
Palmitoylation	C322, C323		C341	
Simulations	Sim 1: 0.89 μs Sim 2: 0.90 μs Sim 3^{t}: 1.00 μs	Sim 1: 0.98 μs Sim 2: 0.98 μs Sim 3^{t}: 1.00 μs Sim 4^{t}: 1.00 μs Sim 5^{t}: 1.00 μs	Sim 1: 1.00 μs Sim 2: 1.00 μs	Sim 1: 1.00 μs Sim 2: 1.00 μs

Simulations marked with a superscript 't' indicate truncation of a crystallographically resolved ICL3.

ARTICLE

https://doi.org/10.1038/s41586-018-0236-6

Structure of the adenosine-bound human adenosine A_1 receptor–G_i complex

Christopher J. Draper-Joyce[1,6], Maryam Khoshouei[2,3,6], David M. Thal[1], Yi-Lynn Liang[1], Anh T. N. Nguyen[1], Sebastian G. B. Furness[1], Hariprasad Venugopal[4], Jo-Anne Baltos[1], Jürgen M. Plitzko[2], Radostin Danev[2], Wolfgang Baumeister[2], Lauren T. May[1], Denise Wootten[1,5], Patrick M. Sexton[1,5]*, Alisa Glukhova[1]* & Arthur Christopoulos[1]*

The class A adenosine A_1 receptor (A_1R) is a G-protein-coupled receptor that preferentially couples to inhibitory $G_{i/o}$ heterotrimeric G proteins, has been implicated in numerous diseases, yet remains poorly targeted. Here we report the 3.6 Å structure of the human A_1R in complex with adenosine and heterotrimeric G_{i2} protein determined by Volta phase plate cryo-electron microscopy. Compared to inactive A_1R, there is contraction at the extracellular surface in the orthosteric binding site mediated via movement of transmembrane domains 1 and 2. At the intracellular surface, the G protein engages the A_1R primarily via amino acids in the C terminus of the $G\alpha_i$ α5-helix, concomitant with a 10.5 Å outward movement of the A_1R transmembrane domain 6. Comparison with the agonist-bound β_2 adrenergic receptor–G_s-protein complex reveals distinct orientations for each G-protein subtype upon engagement with its receptor. This active A_1R structure provides molecular insights into receptor and G-protein selectivity.

Adenosine (ADO) receptors comprise four subtypes within the class A G-protein-coupled receptor (GPCR) superfamily that mediate the actions of the purine nucleoside, ADO[1]. Activation of the A_1R is therapeutically desirable for ischaemia-reperfusion injury, atrial fibrillation, neuropathic pain and others[1]. Although ADO is used clinically to treat supraventricular tachycardia, the development of A_1R-selective agonists for a broader range of disorders has thus far failed, primarily owing to dose-limiting on-target adverse effects[2]. Alternative approaches are thus necessary for improved A_1R drug action, with studies focusing on the potential for greater A_1R selectivity through targeting allosteric sites, or via development of A_1R conformational state-selective biased agonists that can promote beneficial signalling while sparing pathways mediating on-target adverse effects[3,4].

One area for the development of selective A_1R-targeting drugs is the use of structure-based approaches that leverage advances in GPCR structural biology. Indeed, inactive-state, antagonist-bound, structures of the A_1R were solved using X-ray crystallography[5,6]. However, these studies required modification of the A_1R via thermostabilizing mutations and/or fusion proteins, and cannot inform on mechanisms underlying agonist binding, A_1R activation and G-protein interaction. These features are necessary for the rational design of selective A_1R activators, biased agonists or positive allosteric modulators. An alternative approach to overcoming the current dearth of active-state, G-protein-bound, GPCRs is the use of single-particle cryo-electron microscopy (cryo-EM)[7–9]. The promise of cryo-EM in yielding active-state GPCR complexes has recently been realized through the solution of several receptor structures bound to agonists and the heterotrimeric G_s protein, complementing the only crystal structure so far, to our knowledge, of an agonist-bound GPCR–G-protein complex, that of the $\beta_{2A}R$ complexed to G_s protein[7–10]. However, the A_1R preferentially couples to the $G_{i/o}$ family of G-proteins[1]. Indeed, of the more than 800 human GPCRs, most preferentially couple to $G_{i/o}$ proteins. The $G_{i/o}$ family has four members, which are the most abundantly expressed G proteins throughout the body[11]. $G_{i/o}$ protein activation is typically associated with inhibition of adenylate cyclase, resulting in reduced cAMP accumulation, but they also regulate numerous effectors including enzymes, ion channels and small GTPases. Based predominantly on the high expression of both G_{i2} proteins and A_1Rs in brain and, albeit to a lesser degree, in cardiac tissues (both major organs for A_1R therapies), we chose to focus on G_{i2} as a transducer for the A_1R. Here we report the first, to our knowledge, structure of a GPCR coupled to a heterotrimeric G_i protein, specifically the A_1R–G_{i2} complex bound to its endogenous agonist, ADO, solved using Volta phase-plate (VPP) cryo-EM.

Solving the A_1R–G_{i2} complex

To facilitate complex formation, A_1R and G_{i2} were expressed separately in HighFive insect cells and combined after solubilizing in lauryl maltose-neopentyl glycol (LMNG) and cholesteryl hemisuccinate (CHS) with addition of apyrase and ADO (Extended Data Fig. 1). Stabilization of the A_1R–G_{i2} complex was achieved by introducing four $G\alpha_{i2}$ subunit mutations that alter nucleotide binding and affinity for $G\beta\gamma$[8]. This dominant-negative G_{i2} (DNG_{i2}) was sufficient to enable formation of a stable interaction with the receptor while insensitive to GTP (Extended Data Figs. 1 and 2c). The antagonist dipropylcyclopentylxanthine (DPCPX) displayed similar affinities for the A_1R whether alone or in the presence of wild-type $G\alpha_{i2}$ or DNG_{i2} (Extended Data Fig. 2a), whereas ADO displayed biphasic binding curves with a similar dispersion of high and low affinity states in the presence, but not absence, of either wild-type $G\alpha_{i2}$ or DNG_{i2} (Extended Data Fig. 2b); a characteristic feature of agonist binding to many GPCRs[12]. By contrast, agonist-mediated [^{35}S]GTPγS binding to activated Gα subunits was only observed upon combination of the A_1R with wild-type $G\alpha_{i2}$ (Extended Data Fig. 2c).

The A_1R–G_{i2} complex in LMNG detergent micelles was visualized using a Titan Krios microscope equipped with a VPP. After imaging and initial 2D classification (Extended Data Fig. 3a, b), 3D classification yielded a final map at a nominal resolution of 3.6 Å reconstructed from 263,321 particle projections (Fig. 1, Extended Data Fig. 3c, Extended

[1]Drug Discovery Biology and Department of Pharmacology, Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, Victoria, Australia. [2]Department of Molecular Structural Biology, Max Planck Institute of Biochemistry, Martinsried, Germany. [3]Novartis Institutes for Biomedical Research, Novartis Pharma AG, Basel, Switzerland. [4]Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria, Australia. [5]School of Pharmacy, Fudan University, Shanghai, China. [6]These authors contributed equally: Christopher J. Draper-Joyce, Maryam Khoshouei. *e-mail: patrick.sexton@monash.edu; alisa.glukhova@monash.edu; arthur.christopoulos@monash.edu

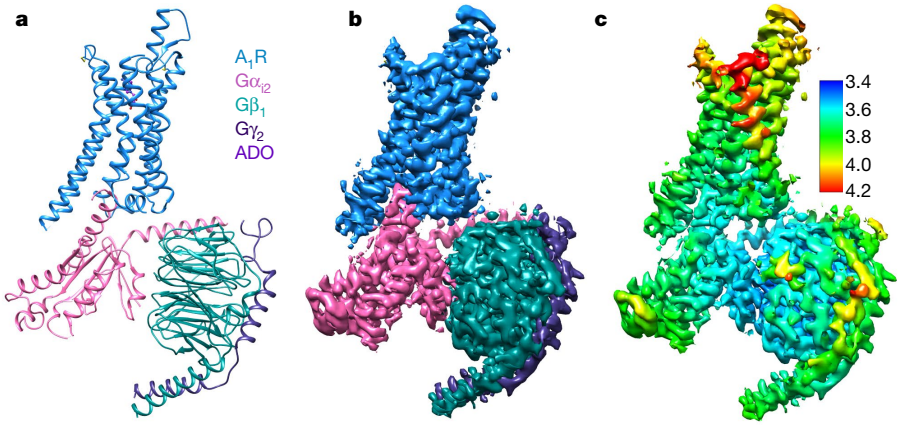


Fig. 1 | The ADO–A_1R–G_{i2} cryo-EM structure. a, Structure determined after refinement in the cryo-EM map (A_1R, blue; ADO, purple; heterotrimeric G_{i2}, pink, cyan and dark purple for α, β and γ, respectively). **b**, 3D cryo-EM map, coloured according to protein chains. **c**, 3D cryo-EM map coloured according to local resolution (Å), defined from half-maps in Relion (v.2.1b1).

Data Table 1). The cryo-EM density map exhibited well-resolved side chains, allowing confident rotamer placements for most amino acids (Extended Data Figs. 3d and 4). The A_1R transmembrane domain regions and extracellular and intracellular loops (ECL and ICL, respectively) are well defined, with the exception of 7 residues in ICL3 and the last 25 residues of helix 8. An ADO molecule is observed occupying the orthosteric site, and extra density at N159 indicates the presence of a glycosylation site that has a role in agonist binding; alanine substitution of this residue caused a significant reduction in the affinity of the ADO derivative NECA (1-(6-amino-9*H*-purin-9-yl)-1-deoxy-*N*-ethyl-β-D-ribofuranuronamide) (Extended Data Fig. 2d–g). The Gβ and Gγ subunits are also well resolved, except for their flexible N and C termini. The α-helical domain of $G\alpha_{i2}$ was averaged out in the 2D class averages owing to high flexibility and thus was masked out during map reconstruction, but the Ras-like domain is well ordered except for the flexible switch III region.

Structure of the active-state A_1R

Notably, despite the presence of a different class of G protein relative to all other active GPCR structures solved so far, there seems to be a general conservation in activation mechanisms. Comparison of the inactive A_1R bound to the covalent antagonist, DU172 (4-[3-(8-cyclohexyl-2,6-dioxo-1-propyl-7H-purin-3-yl)propylcarbamoyl]benzenesulfonyl fluoride; Protein Data Bank (PDB) accession 5UEN)[6], and the active A_1R–G_{i2} bound to ADO, reveals a global allosteric transition similar to that observed upon activation of the β_2AR bound to G_s, or multiple class A active state GPCRs stabilized with nanobodies or a 'mini-$G\alpha_s$' (mG_s) protein[10,13–16] (Fig. 2a–c). A key characteristic of the A_1R activating transition is a large outward movement of the intracellular side of transmembrane helix 6 (TM6) by 10.5 Å to accommodate the α5-helix of the $G\alpha_i$ protein (Fig. 2c). However, this is not as pronounced as in the G_s-bound β_2AR[10] (nor compared to the three class B G_s-bound structures[7–9]), but similar to the change observed in the nanobody-stabilized β_2AR[16], M_2 muscarinic acetylcholine receptor (M_2R)[15], μ-opioid receptor (μOR)[14] and κ-opioid receptor (κOR)[17] structures or the mG_s-stabilized A_{2A}R[13]. Moreover, additional fundamental features of receptor activation previously observed in other (non-$G_{i/o}$ preferring) class A GPCRs are also preserved, including rearrangements in conserved class A activation 'microswitches', such as the PIF, NPXXY and DRY motifs[18] (Fig. 2g), are very similar to the changes observed when comparing the active, agonist-bound, β_2AR–G_s complex and inactive carazolol-bound β_2AR[19]. Small rearrangements at the conserved PIF motif (P192$^{5.50}$, I95$^{3.40}$ and F243$^{6.44}$ in the A_1R; superscripts denote Ballesteros–Weinstein numbering) are associated with the movement of TM6. A 4 Å inward displacement of TM7 at the NPXXY motif (N284$^{7.49}$, P285$^{7.50}$ and Y287$^{7.52}$) propagates to the base of helix 8, and is linked to an outward movement of ICL1. The partially formed 'ionic lock', important in maintaining the ground state of the

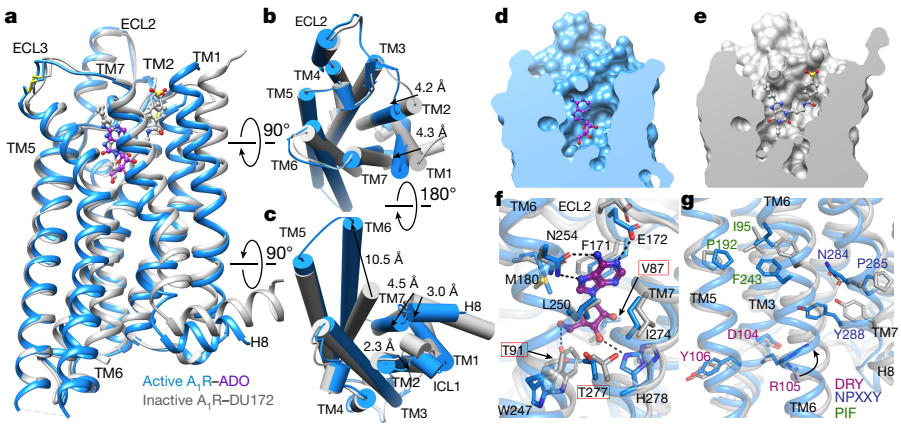


Fig. 2 | Comparison of active and inactive A_1R (PDB code 5UEN) structures. a–**c**, Side (**a**), extracellular (**b**) and cytoplasmic (**c**) views of the receptor (ADO–A_1R–G_{i2} complex in blue; inactive DU172-bound A_1R in grey). **d**, **e**, Active ADO–A_1R (**d**) and inactive DU172–A_1R (**e**) surfaces sliced to show binding site cavity. **f**, Orthosteric binding site of the active A_1R–G_{i2} complex with ADO (purple ball and sticks). 'Toggle switch' W247$^{6.48}$ and residues within 4 Å of ADO are labelled and shown as sticks. Red rectangles highlight rotamer changes upon activation. N, O and S atoms are coloured in blue, red and yellow, respectively. Dashed lines represent hydrogen bonds. **g**, Conserved class A GPCR motifs important for receptor activation (DRY motif, purple; NPXXY motif, blue; PIF motif, green). H8, helix 8.

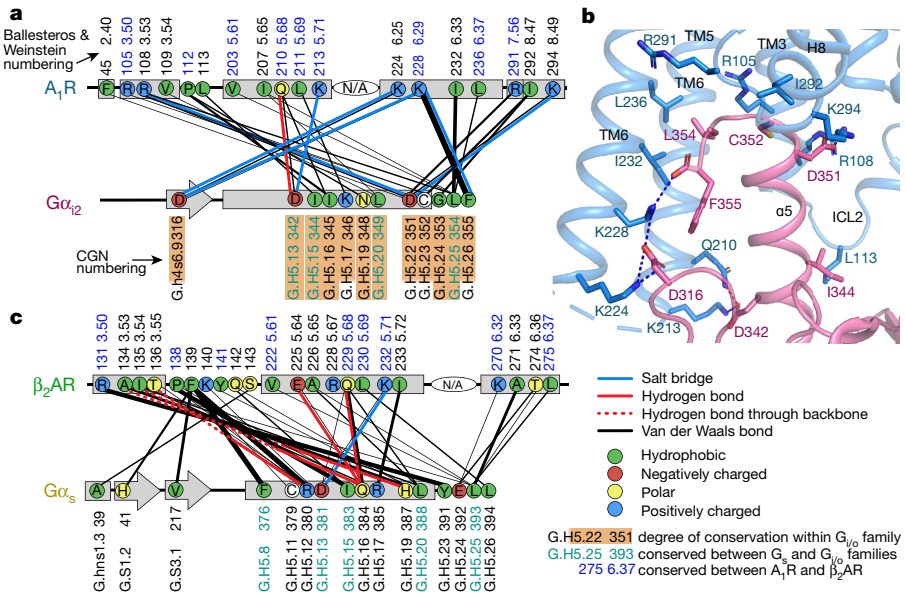


Fig. 3 | Comparison of Gα$_i$ and Gα$_s$ interactions. a, Diagram of the contacts between Gα$_{i2}$ and A_1R. Receptor residues are numbered according to Ballesteros and Weinstein numbering[34]; G-protein residues are numbered according to the CGN scheme[29]. Blue receptor residue numbers indicate conservation between A_1R and β_2AR. Length of orange shading of Gα$_{i2}$ numbers illustrates degree of conservation within the Gα$_{i/o}$ family. **b**, View of key interactions between A_1R (blue) and DNGα$_{i2}$ (pink). Hydrogen bonds and salt bridges are depicted as dashed lines coloured as in panel **a**. **c**, Diagram of the contacts between Gα$_s$ and β_2AR (based on PDB code 3SN6), coloured as in panel **a**.

receptor[20] and observed in the inactive A_1R[6], is broken, and R105$^{3.50}$ of the DRY motif extends towards TM7 to form a lid over the Gα$_{i2}$ α5-helix.

Another key feature of the inactive A_1R is the presence of a large binding cavity that accommodates the orthosteric site[6]. Upon activation, this wide cavity collapses owing to an approximately 4 Å inward movement of the extracellular ends of TM1 and TM2 (Fig. 2b, d, e). Of note, ECL2, which adopts a distinct orientation almost perpendicular to the plane of the membrane, and contributes to the binding of allosteric modulators[3], remains essentially unaltered (Fig. 2a). The collapse of the extracellular cavity is less pronounced when compared to the inactive A_1R bound to the A_1-selective xanthine-based antagonist PSB36 (1-butyl-8-(hexahydro-2,5-methanopentalen-3a(1*H*)-yl)-3,7-dihydro-3-(3-hydroxypropyl)-1*H*-purine-2,6-dione)[5] (Extended Data Fig. 5), probably reflecting that the latter, reversibly binding, antagonist is smaller than the bulky DU172. Nonetheless, a shrinkage of the orthosteric site upon activation is also observed in A_{2A}R[13,21], β_2AR[10,22], M_2R[15,23] and μOR[14,24]. Interestingly, despite notable differences in the ligand-binding sites between the inactive A_1R and A_{2A}R, a comparison of the active states of the two subtypes reveals an almost perfect superimposition of their transmembrane domain regions (root mean squared deviation (r.m.s.d.) value of 1.03) (Extended Data Fig. 6a–c).

The side-chain conformations in the orthosteric site are very similar between the active and inactive A_1Rs. Subtle upward movements of the TM3 backbone (approximately 1.5 Å), and TM7 towards TM6 (1 Å), side-chain movements of W247$^{6.48}$ ('toggle switch') and H278$^{7.43}$, and a change in rotamer conformations of V87$^{3.32}$, T91$^{3.36}$ and T277$^{7.42}$ collectively serve to accommodate the ADO ribose moiety (Fig. 2f). A role for the two threonine residues is consistent with previous studies, whereas mutation of V87A was suggested to have no effect[25,26]. We thus re-investigated the role of V87A and found that, in our hands, the mutation significantly reduced both the affinity and efficacy of NECA (Extended Data Fig. 2d–g). Interestingly, the related (but G_s-preferring) A_{2A}R has previously been solved not only in an active state, bound to NECA and a mG$_s$ protein, but also an 'intermediate' state, bound to either NECA or ADO but in the absence of transducer[13,21,27]. Further comparison of our active A_1R structure with either of the A_{2A}R active or intermediate agonist-bound states revealed similar shifts in TM3 and TM7, and similar changes in W246$^{6.48}$, H278$^{7.43}$, T88$^{3.36}$ and V84$^{3.32}$ after agonist binding[13,21,27] (Extended Data Fig. 6). These complementary rearrangements result in nearly identical orthosteric sites between agonist-bound A_1R and A_{2A}R, suggesting that the propagation and selection of coupling partner must involve conformational changes downstream of this binding site. Other notable interactions between orthosteric site residues and ADO in the A_1R include π-stacking with F171^{ECL2}, hydrogen bonding with N254$^{6.55}$, H278$^{7.43}$ and E172^{ECL2}, and Van der Waals interactions with M180$^{5.38}$ and L250$^{6.51}$ (Fig. 2f, Extended Data Fig. 5f), which are all consistent with previous mutational and computational studies of agonist interactions with the A_1R[25,28].

Structure, coupling and selectivity of the G$_i$ heterotrimer

Despite the recent solution of a number of class A GPCR structures in 'active-like' states stabilized by nanobodies or mG$_s$, there still remains only a single class A GPCR structure bound to a heterotrimeric G protein—the β_2AR–G$_s$ complex[10]. As also observed with that complex, interactions between G$_i$ and the A_1R are extensive, with a total buried surface area of 1,964 Å^2 (958 Å^2 at the A_1R and 1,006 Å^2 at the Gα$_{i2}$). Another similarity between the two GPCR–G-protein complexes is that the G protein interacts primarily via the Gα Ras-like domain. An alignment of both receptors (Extended Data Fig. 7a) and G proteins (Extended Data Fig. 7b), followed by a comparison of key interaction points between the α5-helix of each Gα subunit with its respective receptor, highlighted two distinct clusters (Fig. 3; Extended Data Fig. 4b, c). The first region is common to both Gα$_s$ and Gα$_i$, encompassing approximately residues 12–20 of each α5-helix (H5.8–H5.20, common Gα numbering (CGN) system[29]), possibly highlighting the broad importance of this region in providing strong contacts for receptor binding. However, for the β_2AR–G$_s$ complex, there are more interactions within this region of α5 with TM3 and TM5 of the receptor that consist of both polar and non-polar contacts. By contrast, there are fewer interactions for the A_1R–G$_i$ complex and only D342 (G.H5.13) of the Gα$_i$ subunit forms a hydrogen bond with Q210$^{5.68}$ and a salt bridge with K213$^{5.71}$, with the rest of the cluster making nonpolar interactions. The second key interaction region comprises the five C-terminal α5-helix amino acid residues that provide more receptor interactions for the Gα$_i$ than the Gα$_s$ helix. These last five amino acids provide the strongest set of interactions for A_1R–G$_{i2}$ (Fig. 3a, b), including inter-

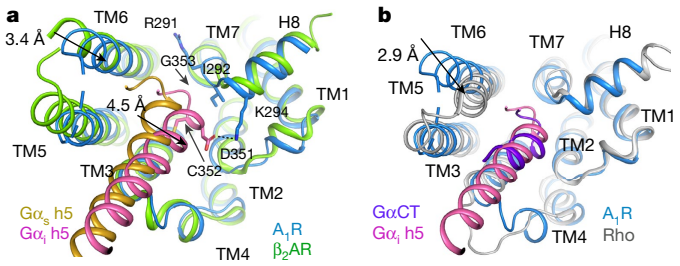


Fig. 4 | Orientation of Gα subunits varies between GPCR–G-protein complexes. a, Cytoplasmic view shows α5-helix orientation difference between G_i and G_s. Interacting residues of the α5-helix (h5), TM7 and helix 8 are shown as sticks. The salt bridge is shown as a dashed line. **b**, Cytoplasmic view reveals a similar orientation of the $G\alpha_{i2}$ α5-helix and a synthetic peptide derived from the Gα subunit of transducin (GαCT, purple; PDB code 3PQR) relative to active A_1R and active Rho (in grey), respectively. The small inward movement of TM6 probably reflects a lack of additional contacts for Rho beyond the 11-mer GαCT fragment, whereas the A_1R engages the entire G_{i2} heterotrimer.

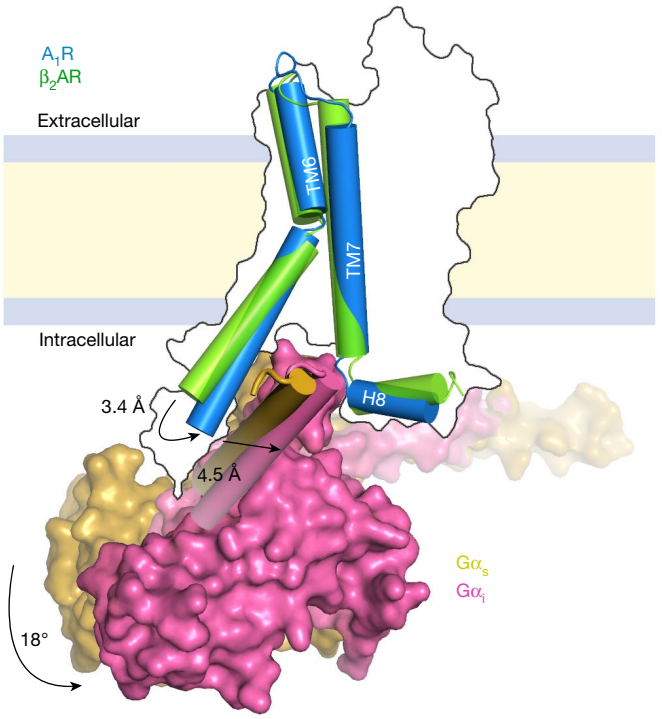


Fig. 5 | Schematic summarizing the key translational and rotational movements contributing to differences in G-protein coupling between A_1R–G_{i2} with β_2AR–G_s. The receptor outline is that of the A_1R as a reference. Key receptor transmembrane domains and the Gα protein C-terminal α5-helices are shown as cylindrical helixes (β_2AR, green; A_1R, blue); the rest of the Gα proteins are represented as contoured surfaces with $G\alpha_{i2}$ in pink and $G\alpha_s$ in gold.

actions with TM2, TM3, TM5–TM7 and helix 8. The carboxyl group of F355 (GH5.26) forms a salt bridge with K228$^{6.29}$ (usually K/R in G_i-coupled receptors[29]). L354 (GH5.25) extends from Gα into a hydrophobic pocket lined with A_1R I232$^{6.33}$, L236$^{6.37}$, V203$^{5.61}$ and I207$^{5.65}$. Residue D351 (GH5.22, D in $G\alpha_{i1/2}$ and E in $G\alpha_{i3}$) forms a salt bridge with R108$^{3.53}$ and K294$^{8.49}$. Notably, studies of chimeric G proteins have suggested that these five C-terminal α5-helix residues are generally sufficient for selective GPCR–G-protein coupling[30], although other residues are also involved[31]. Nonetheless, it is evident that this region makes a very different set of interactions with the A_1R compared to the same region in the β_2AR–Gs complex α5-helix, the latter characterized by weaker Van der Waals interactions (Fig. 3c). In addition to the α5-helical residues, D316 of the α4–β6-loop of $G\alpha_{i2}$ stabilizes the interaction with the A_1R via a salt bridge with K224$^{6.25}$ and K228$^{6.29}$ (Fig. 3a, c).

Although both $G\alpha_i$ and $G\alpha_s$ interact with TM3, TM5, ICL1 and TM6 of their respective GPCRs (Extended Data Fig. 8), $G\alpha_i$ makes additional contacts with the cytoplasmic end of TM7 (R291$^{7.56}$) and helix 8 (I292$^{8.47}$ and R294$^{8.49}$), which are absent in the β_2AR–G_s structure (Fig. 4a). In addition, extra density near the unmodelled flexible ICL3 of the A_1R suggests the potential for additional interactions between the receptor and the $G\alpha_i$ β6-sheet. Notably, A_1R L113^{ICL2} does not contribute substantially to $G\alpha_i$ binding, in contrast to the equivalent residue, F139^{ICL2}, in β_2AR–G_s. Indeed, L113^{ICL2} forms only a single Van der Waals bond with I344 (GH5.15), whereas β_2AR F139^{ICL2} binds in an extensive hydrophobic pocket formed by the $G\alpha_s$ α5-helix, β2–β3-loop and β1-strands (Fig. 3a–c). As in the β_2AR–G_s complex, ICL1 and helix 8 of the A_1R are in close proximity to the Gβ subunit (Extended Data Fig. 9a). The presence of multiple polar and charged amino acids at this interface suggests potential interactions; however, the side-chain density was insufficient for confident modelling.

There have been numerous studies identifying a role for the Gα subunit α5-helix C terminus in determining G-protein selectivity for GPCRs[31], but less is known about receptor determinants that govern GPCR selectivity for G proteins[29]. A receptor-based alignment of the A_1R–G_{i2} and β_2AR–G_s complexes revealed that each receptor engages its G protein in a different orientation (Fig. 4a; Extended Data Fig. 8); compared to β_2AR–G_s, G_i in the A_1R–G_{i2} complex is translated by approximately 4.5 Å relative to the receptor along the TM5–TM1 vector. This translation arises from a difference in the position of the α5-helix relative to the receptor, leading to a movement of the rest of $G\alpha_i$ coupled with movements of Gβ and Gγ subunits. There are two likely mechanisms contributing to the observed translation: conformational differences in the Gα subunits and distinct receptor conformations. With respect to the first mechanism, a comparison of the nucleotide-free, receptor-bound conformations of the G_i and G_s α-subunits revealed that while, overall, they adopt a similar backbone conformation (r.m.s.d. value of 1.34 Å), there are differences in the flexible loop regions and the α5-helix conformation (Extended Data Fig. 9b), which arise as a consequence of the last two turns of the $G\alpha_s$ α5-helix bending towards the α4–β6-loop. This results in a 3.5 Å shift in positioning of the G.H5.23 C_α residue (Extended Data Fig. 9c). However, the bend in the $G\alpha_s$ α5-helix was not observed in cryo-EM structures[7–9]. The second mechanism that may lead to distinct orientations of $G\alpha_{i2}$ relative to $G\alpha_s$ in complex with their receptors is related to conformational differences in the receptors themselves, particularly TM6. The A_1R structure displays a smaller outward movement of TM6 relative to β_2AR–G_s[10] (10.5 Å versus 14 Å, respectively), with the α5-helix closer to TM7 and helix 8 that results in a difference in receptor interaction angle (approximately 18°) (Fig. 4a). This translation brings D351 (G.H5.22) within a salt-bridge distance of K294$^{8.49}$, and both C352 (G.H5.23) and G353 (G.H5.24) within Van der Waals radius of R291$^{7.56}$ and I292$^{8.47}$. A comparison of A_1R–G_{i2} with metarhodopsin II (Rho), the active state of rhodopsin, bound to a C-terminal fragment of transducin (GαCT) reveals a similar location of the two domains for $G\alpha_i$ and transducin ($G\alpha_t$) (Fig. 4b), consistent with the fact that $G\alpha_t$ is most closely related to G_i proteins[32]. The key global conformational differences between the β_2AR–G_s and A_1R–G_{i2} complexes are summarized in Fig. 5. Although it remains to be determined whether G-protein selectivity is mediated predominantly by specific receptor residue contacts as opposed to global conformational rearrangements shaping intracellular 'pocket complementarity'[29,33], a key role for the degree of TM6 tilt is a possibility. However, this should be interpreted with caution because the distinct differences in the movement of TM6 between the β_2AR–G_s and A_1R–G_{i2} structures and translation of the G protein could actually manifest as a result of three different underlying mechanisms: (i) true differences in receptor activation independent of G-protein coupling; this is not readily apparent upon comparison of

the A_1R and $A_{2A}R$ structures (Extended Data Fig. 6); (ii) the signalling state of the solved receptor–G-protein complex being influenced by the nanobody (Nb35) present in the β_2AR–G_s or the dominant-negative mutations present in the $G\alpha_{i2}$ protein in our A_1R–G_{i2} structure; or (iii) a difference in the general activation mechanism for $G_{i/o}$-coupled compared to G_s-coupled GPCRs. This will ultimately require the solution of more GPCR–G-protein structures in as 'native' a state as possible. Nonetheless, a role for the degree of TM6 tilt in G-protein selectivity is consistent with previous molecular dynamics studies on β_2AR complexes with C-terminal $G\alpha_s$- or $G\alpha_i$-derived peptides[33] and this is supported in the available G protein complexed GPCR structures.

In conclusion, this study presents the first, to our knowledge, active-state class A cryo-EM structure, highlighting the utility of this technique to determine new GPCR structural information. The structure of the ADO–A_1R–G_{i2} heterotrimer provides important insights into the activation mechanism of the A_1R in response to its endogenous agonist. It also allows the first comparison between different subtypes of heterotrimeric G proteins bound to activated GPCRs, which may be pivotal in understanding pleiotropic coupling of these receptors. Finally, the findings have broad implications for understanding the structural basis underlying GPCR–G-protein selectivity, and can facilitate rational, structure-based, approaches for the design of subtype-selective A_1R ligands as new therapeutic agents.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0236-6.

Received: 6 February 2018; Accepted: 16 May 2018;
Published online: 20 June 2018

1. Fredholm, B. B., IJzerman, A. P., Jacobson, K. A., Linden, J. & Müller, C. E. International Union of Basic and Clinical Pharmacology. LXXXI. Nomenclature and classification of adenosine receptors—an update. *Pharmacol. Rev.* **63**, 1–34 (2011).
2. Jacobson, K. A. & Gao, Z.-G. Adenosine receptors as therapeutic targets. *Nat. Rev. Drug Discov.* **5**, 247–264 (2006).
3. Nguyen, A. T. N. et al. Role of the second extracellular loop of the adenosine A_1 receptor on allosteric modulator binding, signaling, and cooperativity. *Mol. Pharmacol.* **90**, 715–725 (2016).
4. Valant, C. et al. Separation of on-target efficacy from adverse effects through rational design of a bitopic adenosine receptor agonist. *Proc. Natl Acad. Sci. USA* **111**, 4614–4619 (2014).
5. Cheng, R. K. Y. et al. Structures of human A_1 and A_{2A} adenosine receptors with xanthines reveal determinants of selectivity. *Structure* **25**, 1275–1285.e4 (2017).
6. Glukhova, A. et al. Structure of the adenosine A_1 receptor reveals the basis for subtype selectivity. *Cell* **168**, 867–877.e13 (2017).
7. Liang, Y.-L. et al. Phase-plate cryo-EM structure of a class B GPCR–G-protein complex. *Nature* **546**, 118–123 (2017).
8. Liang, Y.-L. et al. Phase-plate cryo-EM structure of a biased agonist-bound human GLP-1 receptor–G_s complex. *Nature* **555**, 121–125 (2018).
9. Zhang, Y. et al. Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature* **546**, 248–253 (2017).
10. Rasmussen, S. G. F. et al. Crystal structure of the β_2 adrenergic receptor–Gs protein complex. *Nature* **477**, 549–555 (2011).
11. Alexander, S. P. et al. The Concise Guide To Pharmacology 2017/18: G protein-coupled receptors. *Br. J. Pharmacol.* **174** (Suppl. 1), S17–S129 (2017).
12. De Lean, A., Stadel, J. M. & Lefkowitz, R. J. A ternary complex model explains the agonist-specific binding properties of the adenylate cyclase-coupled beta-adrenergic receptor. *J. Biol. Chem.* **255**, 7108–7117 (1980).
13. Carpenter, B., Nehmé, R., Warne, T., Leslie, A. G. W. & Tate, C. G. Structure of the adenosine A_{2A} receptor bound to an engineered G protein. *Nature* **536**, 104–107 (2016).
14. Huang, W. et al. Structural insights into μ-opioid receptor activation. *Nature* **524**, 315–321 (2015).
15. Kruse, A. C. et al. Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* **504**, 101–106 (2013).
16. Rasmussen, S. G. F. et al. Structure of a nanobody-stabilized active state of the β_2 adrenoceptor. *Nature* **469**, 175–180 (2011).
17. Che, T. et al. Structure of the nanobody-stabilized active state of the kappa opioid receptor. *Cell* **172**, 55–67.e15 (2018).
18. Erlandson, S. C., McMahon, C. & Kruse, A. C. Structural basis for G protein-coupled receptor signaling. *Annu. Rev. Biophys.* **47**, 1–18 (2018).
19. Rosenbaum, D. M. et al. GPCR engineering yields high-resolution structural insights into β_2-adrenergic receptor function. *Science* **318**, 1266–1273 (2007).
20. Savinainen, J. R., Saario, S. M., Niemi, R., Järvinen, T. & Laitinen, J. T. An optimized approach to study endocannabinoid signaling: evidence against constitutive activity of rat brain adenosine A_1 and cannabinoid CB_1 receptors. *Br. J. Pharmacol.* **140**, 1451–1459 (2003).
21. Liu, W. et al. Structural basis for allosteric regulation of GPCRs by sodium ions. *Science* **337**, 232–236 (2012).
22. Rasmussen, S. G. F. et al. Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383–387 (2007).
23. Haga, K. et al. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* **482**, 547–551 (2012).
24. Manglik, A. et al. Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature* **485**, 321–326 (2012).
25. Rivkees, S. A., Barbhaiya, H. & IJzerman, A. P. Identification of the adenine binding site of the human A_1 adenosine receptor. *J. Biol. Chem.* **274**, 3617–3621 (1999).
26. Townsend-Nicholson, A. & Schofield, P. R. A threonine residue in the seventh transmembrane domain of the human A_1 adenosine receptor mediates specific agonist binding. *J. Biol. Chem.* **269**, 2373–2376 (1994)
27. Lebon, G. et al. Agonist-bound adenosine A_{2A} receptor structures reveal common features of GPCR activation. *Nature* **474**, 521–525 (2011).
28. Nguyen, A. T. N. et al. Extracellular loop 2 of the adenosine A_1 receptor has a key role in orthosteric ligand affinity and agonist efficacy. *Mol. Pharmacol.* **90**, 703–714 (2016).
29. Flock, T. et al. Universal allosteric mechanism for Gα activation by GPCRs. *Nature* **524**, 173–179 (2015).
30. Stewart, G. D. et al. Determination of adenosine A_1 receptor agonist and antagonist pharmacology using *Saccharomyces cerevisiae*: implications for ligand screening and functional selectivity. *J. Pharmacol. Exp. Ther.* **331**, 277–286 (2009).
31. Oldham, W. M. & Hamm, H. E. Heterotrimeric G protein activation by G-protein-coupled receptors. *Nat. Rev. Mol. Cell Biol.* **9**, 60–71 (2008).
32. Choe, H.-W. et al. Crystal structure of metarhodopsin II. *Nature* **471**, 651–655 (2011).
33. Rose, A. S. et al. Position of transmembrane helix 6 determines receptor G protein coupling specificity. *J. Am. Chem. Soc.* **136**, 11244–11247 (2014).
34. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci* **25**, 366–428 (1995).

Acknowledgements This work was supported by the Monash University Ramaciotti Centre for Cryo-Electron Microscopy, National Health and Medical Research Council of Australia (NHMRC) project grant APP1145420 and NHMRC program grant APP1055134. A.C., P.M.S. and D.W. are NHMRC Senior Principal Research, Principal Research and Career Development Fellows, respectively. A.G. and D.M.T. are Australian Research Council Discovery Early Career Research Fellows. L.T.M. is an Australian Heart Foundation Future Leaders Fellow.

Reviewer information *Nature* thanks D. Wacker and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.J.D.-J. developed the expression and purification strategy, performed virus production, insect cell expression, purification, and membrane-based pharmacological assays, negative-stain electron microscopy data acquisition/analysis and prepared samples for cryo-EM; M.K. performed sample plunging for cryo-EM, phase-plate imaging and data collection, electron microscopy data processing and analysis; D.M.T. developed the expression and purification strategy, assisted with biochemistry, structure refinement and validation and model interpretation; Y.-L.L. and S.G.B.F. developed the strategy to generate the dominant-negative $G\alpha_{i2}$; R.D. and W.B. organized and developed the Volta phase-plate cryo-EM data acquisition strategy; H.V. organized microscopy time and provided oversight of image acquisition within the Monash EM facility; J.M.P. provided advice on microscope setup for phase-plate imaging and EM facility access within the Max Planck Institute; A.T.N.N. and J.A.B. performed whole-cell pharmacological assays; L.T.M. supervised whole-cell pharmacological assays; C.J.D.-J., A.T.N.N., J.A.B. and L.T.M. performed data analysis; M.K., D.M.T., Y.-L.L., S.G.B.F., L.T.M., D.W. and P.M.S. assisted with data interpretation and preparation of the manuscript; A.G. developed the expression and purification strategy, performed negative-stain electron microscopy, cryo-EM sample preparation, model building, refinement and validation; C.J.D.-J., A.G. and A.C. wrote the manuscript; P.M.S., A.G. and A.C. supervised the project.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0236-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0236-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to P.M.S., A.G. or A.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Constructs. Wild-type human A_1R was modified to include an N-terminal Flag tag epitope and a C-terminal 8× histidine tag; both tags are removable by 3C protease cleavage (Extended Data Fig. 1a). A DNGα_{i2} construct was generated by site-directed mutagenesis to incorporate mutations that alter nucleotide binding (S47N, G204A and A327S) and a mutation (E246A) that improves the dominant-negative effect by weakening a salt bridge that helps to stabilize the nucleotide-bound conformation[8,35–38]. These constructs were generated in insect cell expression vectors.

Insect cell expression. A_1R, human DNGα_{i2}, His_6-tagged human Gβ_1 and Gγ_2 were expressed in HighFive insect cells (Expression Systems) using baculovirus. Cell cultures were grown in ESF 921 serum-free media (Expression System) to a density of 4 million cells per ml and then infected with either A_1R baculovirus or both Gα_{i2} and G$\beta_1\gamma_2$ baculovirus, at a ratio of 1:1. Cultures were grown at 27 °C and collected by centrifugation 60 h after infection. Cells were snap frozen and stored at −80 °C for later use.

Complex purification. Cells from either A_1R or heterotrimeric G_i expression were solubilized separately in 20 mM HEPES pH 7.4, 100 mM NaCl, 5 mM $MgCl_2$, 5 mM $CaCl_2$, 0.5% (w/v) lauryl maltose neopentyl glycol (LMNG, Anatrace), 0.01% (w/v) cholesteryl hemisuccinate (CHS, Anatrace) supplemented with cOmplete Protease Inhibitor Cocktail tables (Roche). Complex formation was initiated combining solubilized A_1R and heterotrimeric G_i and by addition of 1 mM ADO (Sigma) and apyrase (25 mU ml^{-1}, NEB); followed by 2 h incubation at 4 °C. Insoluble material was removed by centrifugation at 30,000*g* for 30 min and the solubilized complex was immobilized by batch binding to M1 anti-Flag affinity resin in the presence of 5 mM $CaCl_2$. The resin was packed into a glass column and washed with 20 column volumes of 20 mM HEPES pH 7.4, 100 mM NaCl, 5 mM $MgCl_2$, 5 mM $CaCl_2$, 1mM ADO, 0.01% (w/v) LMNG and 0.001% (w/v) CHS before bound material was eluted in buffer containing 10 mM EGTA and 0.1 mg ml^{-1} Flag peptide. The complex was concentrated using an Amicon Ultra Centrifugal Filter (MWCO 100 kDa) and subjected to size-exclusion chromatography on a Superdex 200 Increase 10/300 column (GE Healthcare) pre-equilibrated with 20 mM HEPES pH 7.4, 100 mM NaCl, 5 mM $MgCl_2$, 1 mM ADO, 0.01% (w/v) and 0.001% (w/v) CHS to separate complex from contaminants. Eluted fractions consisting of receptor and G-protein complex were pooled and concentrated. Final yield of purified complex was approximately 0.2 mg l^{-1} insect cell culture.

SDS–PAGE and western blot analysis. Samples collected from each purification step were analysed by SDS–PAGE and western blot. For SDS–PAGE, precast gradient TGX gels (Bio-Rad) were used. Gels were either stained by Instant Blue (Expedeon) or immediately transferred to PVDF membrane (Bio-Rad) at 100 V for 1 h. The proteins on the PVDF membrane were probed first with the primary mouse anti-His antibody (34660, QIAGEN), followed by washing and incubation with secondary anti-mouse antibody (680RD). The membranes were again washed and incubated with an AlexaFluor488 conjugated Gα_{i2} mouse monoclonal antibody (sc13534, Santa Cruz) against Gα_{i2}. Bands were imaged using a Typhoon multimode imager (GE Healthcare Life Sciences).

Preparation of vitrified specimen. For cryo-EM, purified A_1R–G_{i2} complex was diluted to 1.3 mg ml^{-1} with 20 mM HEPES 7.4, 100 mM NaCl, 5 mM $MgCl_2$, 1 mM ADO. EM grids (Quantifoil, 200 mesh copper R1.2/1.3) were glow discharged for 30 s using Harrick plasma cleaner (Harrick). 4 μl sample was applied on the grid in the Vitrobot chamber (FEI Vitrobot Mark IV). The chamber of Vitrobot was set to 100% humidity at 4 °C. The sample was blotted for 4.5 s with blot force of 20 and then plunged into propane-ethane mixture (37% ethane and 63% propane).

Data acquisition. Datasets were collected on an FEI Titan Krios microscope operated at 300 kV (FEI) equipped with a Gatan Quantum energy filter, a Gatan K2 summit direct electron camera (Gatan) and a Volta phase-plate[39–42] (FEI). Movies were taken in EFTEM nanoprobe mode, with 50 μm C2 aperture, at a calibrated magnification of 47,170 corresponding to a magnified pixel size of 1.06 Å. Each movie comprises 50 sub frames with a total dose of 50 e^- $Å^{-2}$, exposure time was 12.5 s with the dose rate of 4.5 e^- $pixels^{-1}$ s^{-1} on the detector. Data acquisition was done using SerialEM software at −500 nm defocus[43].

Data processing. A total of 3,220 movies were collected and subjected for motion correction using motioncor2[44]. The movies were collected after installation of a new VPP. Each position on the VPP was used for 24 h. CTF estimation was done using Gctf software[45] on non-dose-weighted micrographs. The particles were picked using gautomatch (developed by K. Zhang, MRC Laboratory of Molecular Biology, Cambridge, UK, http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/). An initial model was made using EMAN2[46] based on automatically picked few micrographs and using the common-line approach. The particles were extracted in RELION 2.01b1[47] using a box size of 180 pixels. Next, 831,602 picked particles were subjected to two rounds of 3D classification with three classes. After selecting the best-looking class with 263,321 particles, 3D auto-refinement was done in RELION 2.01b1. The final map was sharpened with a *B*-factor of −196 $Å^2$. Local resolution was determined using the internal local resolution procedure in Relion, using half-reconstruction as input maps.

Model building. The initial receptor model for A_1R–G_{i2} complex refinement was a chimaera between an active state A_1R homology model, made using the SWISS-MODEL server[48], based on the A_{2A}–mG_s structure (PDB code 5G53) and the inactive A_1R structure (PDB code 5UEN). The initial model for the receptor-bound Gα_{i2} subunit was a chimaera between a Gα_{i2} homology model, based on the β_2AR-bound heterotrimeric G_s structure (PDB code 3SN6), and a homology model based on the GDP-bound dominant-negative G_{i1} heterotrimer structure (PDB code 5TDH). Gβ and Gγ models were taken from the β_2AR–G_s structure. Refinement, using the phenix.real_space_refine module in PHENIX[49], was iterated with manual model adjustment and rebuilding in COOT[50]. Restraints for the agonist ADO were generated using the GRADE server, https://www.globalphasing.com (v.12.13). Model validation was performed in MolProbity[51]. For validation of model overfitting to the cryo-EM maps we displaced the model atoms (the same model used for the final refinement step) up to 0.5 Å using 'Shake' in the CCPEM suite[52]. Subsequently, refinement was performed against the sharpened half-map 1 (HM1) (sharpened in Refmac using the full map as the reference[53]) using phenix.real_space_refine module in PHENIX[49] with the same parameters used in the final refinement step. Fourier shell correlation (FSC) curves were generated in the MTRIAGE module[54] in Phenix for: 1) shaken pdb refined against HM1 versus HM1; 2) shaken pdb refined against HM1 versus HM2 (also sharpened in Refmac using the full map as the reference); 3) the final pdb versus the full sharpened map (also sharpened in Refmac using the full map as the reference). Figures were prepared in USCF Chimera[55] or PyMOL Molecular Graphics System, Version 2.0 (Schrödinger, LLC).

Insect cell membrane preparations for [^{3}H]DPCPX and [^{35}S]GTPγS binding. The A_1R complex with wild-type heterotrimeric G_{i2} (A_1R, wild-type Gα_{i2}, G$\beta_1\gamma_2$), DNG$_{i2}$ (A_1R, DNGα_{i2}, G$\beta_1\gamma_2$) or A_1R alone were expressed in HighFive insect cells (Expression systems). Cells were collected approximately 48 h after the viral infection. For crude membrane preparations, cells were resuspended in membrane buffer (20 mM HEPES 7.4, 100 mM $NaCl_2$, 5 mM $MgCl_2$, 1 mM EDTA, with protease inhibitors and benzonase), dounced 20 times, followed by centrifugation (5 min, 350*g*, 4 °C). The pellet was again resuspended in membrane buffer, dounced and clarified by centrifugation at low *g*. Membranes were pelleted by centrifugation (1 h, 30,000*g*, 4 °C), resuspended in the membrane buffer and sonicated. The protein concentration was determined using Bradford reagent (Bio-Rad).

Radioligand competition binding experiments in membranes from insect cells expressing the A_1R. [^{3}H]DPCPX binding was performed in 20 mM HEPES pH 7.4, 100 mM NaCl, 5 mM $MgCl_2$, 0.1% BSA. Membranes were incubated for 1 h at 37 °C with 1 nM of [^{3}H]DPCPX and increasing concentrations of either ADO or cold DPCPX. Membranes were collected on UniFilter GF/C (Whatman) plates using Filtermate 196 harvester (Packard), extensively washed with ice-cold NaCl, dried and dissolved in 50 μl MicroScint-O scintillation cocktail (Packard) and counted using a MicroBeta LumiJET counter (PerkinElmer). Non-specific binding was determined in the presence of 10 μM SLV320. Competition binding curves between [^{3}H]DPCPX and either unlabelled DPCPX or ADO were fitted to one- or two-site competition binding equations in Prism 6.0 (GraphPad).

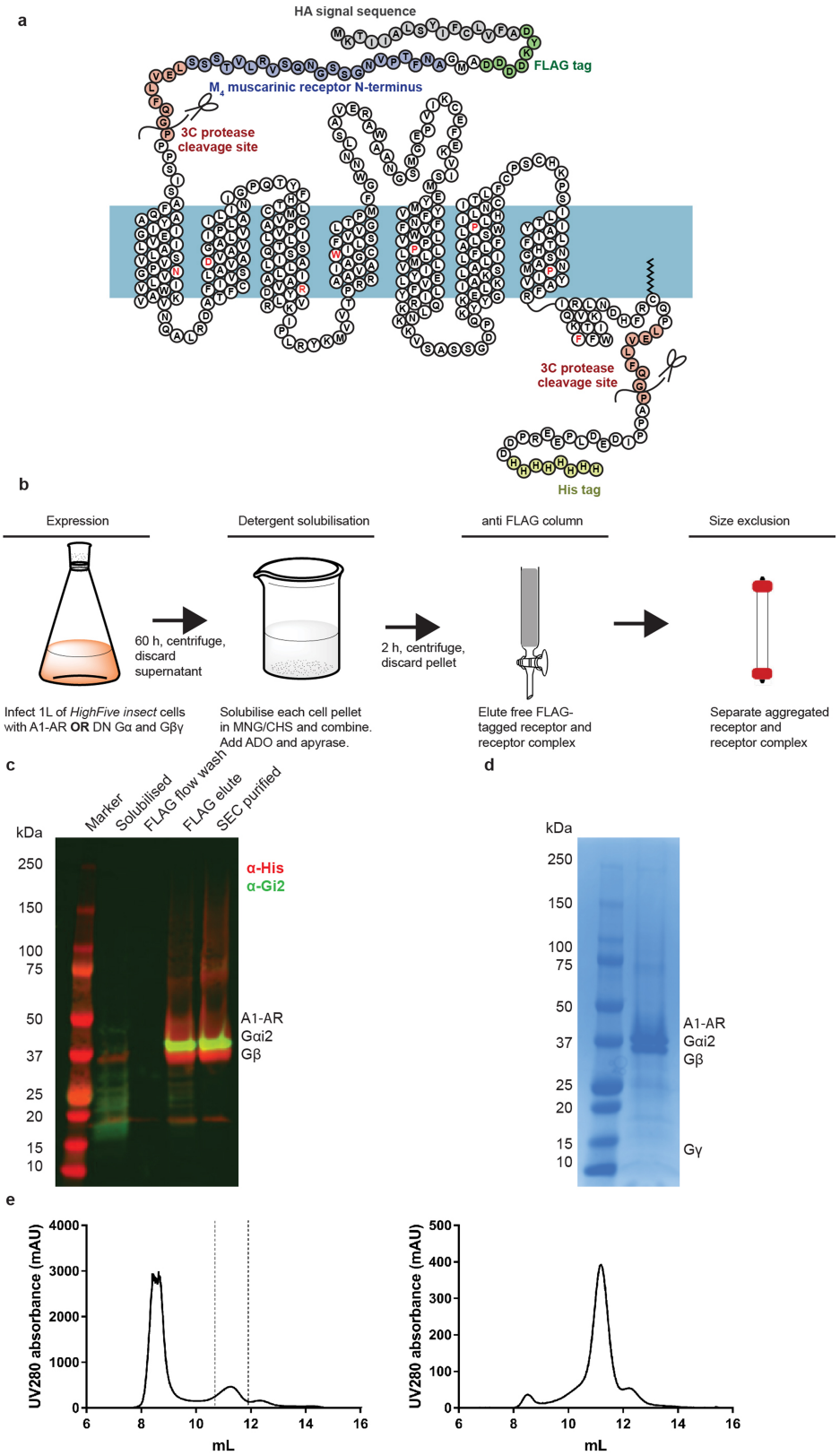
[^{35}S]GTPγS binding in membranes from insect cells expressing the A_1R. Measurement of [^{35}S]GTPγS incorporation to activated G proteins was performed in 20 mM HEPES pH 7.4, 100 mM NaCl, 10 mM $MgCl_2$, 1 mM EDTA, 0.1% BSA, 30 μg ml^{-1} saponin. First, membranes (10 μg per sample) were pre-incubated with 1 μM GDP and increasing concentrations of ADO for 45 min at 22 °C. Reactions were then started by the addition of [^{35}S]GTPγS and ATP to final concentrations of 300 pM and 50 μM, respectively. After 1 h incubation of 30 °C, the reaction was terminated by harvesting the membranes on Whatman UniFilter GF/C plates using Filtermate 196 harvester (Packard). Membranes were extensively washed with ice-cold 50 mM Tris pH 7.6, 10 mM $MgCl_2$, 100 mM NaCl, dried, dissolve in 50 μl MicroScint-O scintillation cocktail (Packard) and counted using a MicroBeta LumiJET counter (PerkinElmer).

Radioligand competition binding and cAMP accumulation assays in whole cells expressing the wild-type or mutant A_1R. Generation of Chinese hamster ovary (CHO) FlpIN cell lines, stably expressing a 3× haemagglutinin (HA)-tagged wild-type human A_1R or alanine substitution mutations (V87A; N159A), [^{3}H]DPCPX competition binding, and inhibition of forskolin-stimulated cAMP accumulation by the agonist, NECA, were all performed as described previously[28]. Generation of the N159A mutation has also been previously described[28]. For V87A, the following oligonucleotides were used to introduce the valine to alanine substitution at position 87; forward: CATGGTTGCCTGTCCGGCCCTCATCCTCACCCAG, reverse: CTGGGTGAGGATGAGGGCCGGACAGGCAACCATG. All other details were as previously described[28]. Cells were routinely tested and confirmed to be free from mycoplasma contamination.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

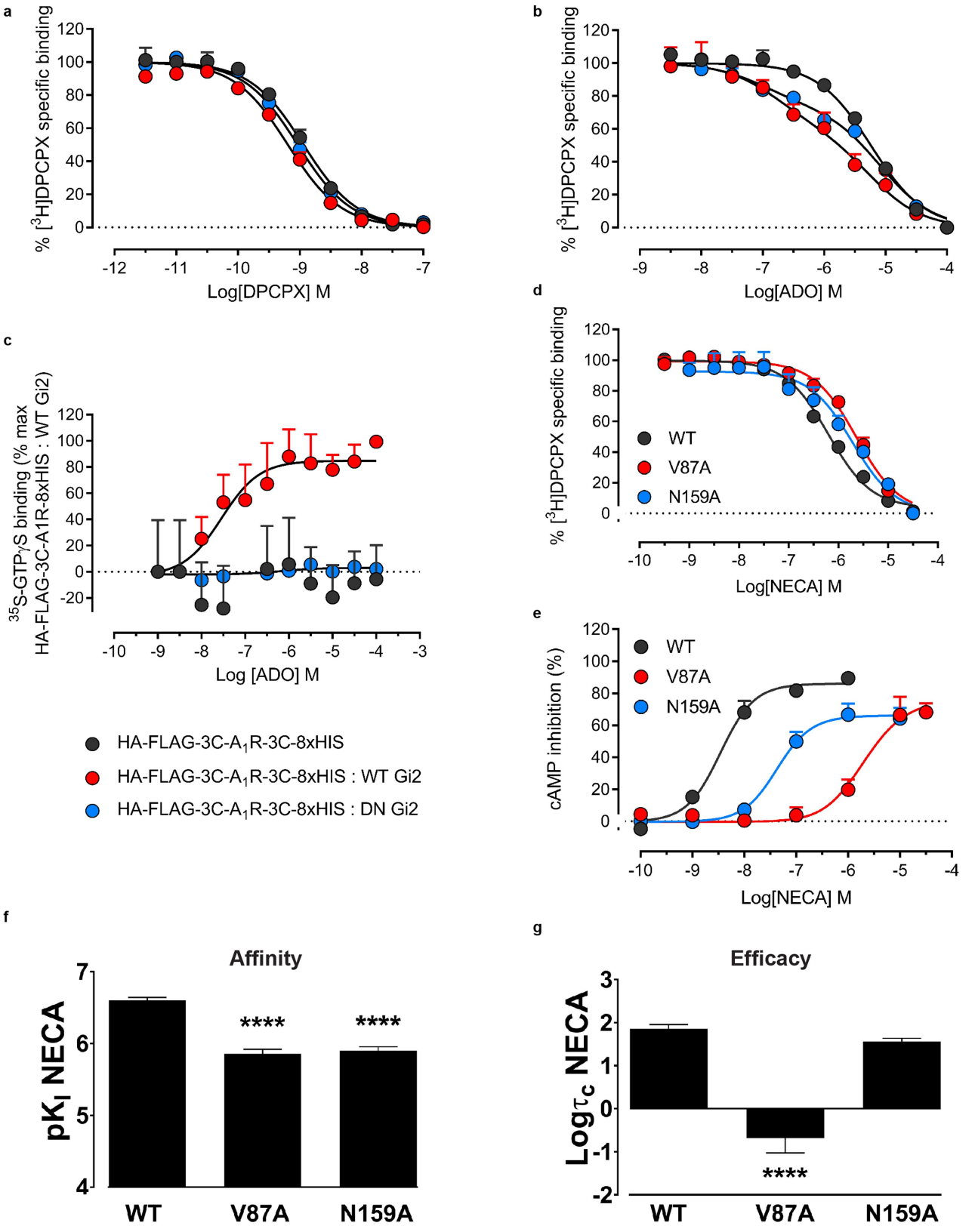
Data availability. All relevant data are available from the corresponding authors and/or included in the manuscript or Supplementary Information. Atomic coordinates and the cryo-EM density map have been deposited in the Protein Data Bank (PDB) under accession number 6D9H and Electron Microscopy Data Bank (EMDB) entry EMD-7835.

35. Cleator, J. H., Mehta, N. D., Kurtz, D. T. & Hildebrandt, J. D. The N54 mutant of $G\alpha_s$ has a conditional dominant negative phenotype which suppresses hormone-stimulated but not basal cAMP levels. *FEBS Lett.* **443**, 205–208 (1999).
36. Iiri, T., Bell, S. M., Baranski, T. J., Fujita, T. & Bourne, H. R. A $G_{s\alpha}$ mutant designed to inhibit receptor signaling through G_s. *Proc. Natl Acad. Sci. USA* **96**, 499–504 (1999).
37. Lee, E., Taussig, R. & Gilman, A. G. The G226A mutant of $G_{s\alpha}$ highlights the requirement for dissociation of G protein subunits. *J. Biol. Chem.* **267**, 1212–1218 (1992).
38. Posner, B. A., Mixon, M. B., Wall, M. A., Sprang, S. R. & Gilman, A. G. The A326S mutant of $G_{i\alpha1}$ as an approximation of the receptor-bound state. *J. Biol. Chem.* **273**, 21752–21758 (1998).
39. Danev, R., Buijsse, B., Khoshouei, M., Plitzko, J. M. & Baumeister, W. Volta potential phase plate for in-focus phase contrast transmission electron microscopy. *Proc. Natl Acad. Sci. USA* **111**, 15635–15640 (2014).
40. Khoshouei, M., Danev, R., Plitzko, J. M. & Baumeister, W. Revisiting the structure of hemoglobin and myoglobin with cryo-electron microscopy. *J. Mol. Biol.* **429**, 2611–2618 (2017).
41. Khoshouei, M. et al. Volta phase plate cryo-EM of the small protein complex Prx3. *Nat. Commun.* **7**, 10534 (2016).
42. Khoshouei, M., Radjainia, M., Baumeister, W. & Danev, R. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nat. Commun.* **8**, 16099 (2017).
43. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
44. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
45. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
46. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
47. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, 19 (2016).
48. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195–201 (2006).
49. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
51. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
52. Burnley, T., Palmer, C. M. & Winn, M. Recent developments in the CCP-EM software suite. *Acta Crystallogr. D* **73**, 469–477 (2017).
53. Brown, A. et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D* **71**, 136–153 (2015).
54. Afonine, P.V. et al. New tools for the analysis and validation of Cryo-EM maps and atomic models. *bioRxiv* https://doi.org/10.1101/279844 (2018).
55. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).



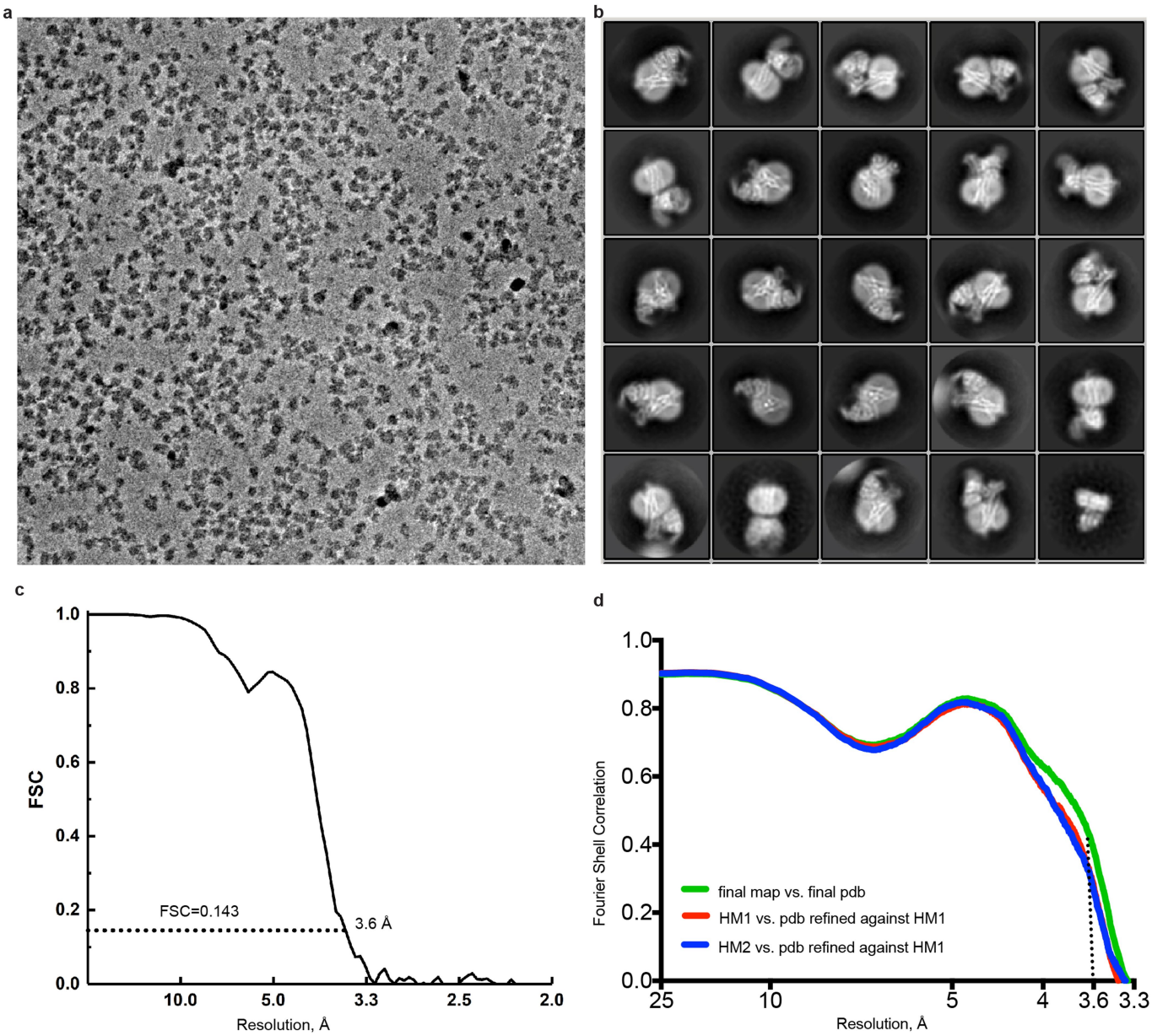
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Expression and purification of the ADO–A_1R–G_{i2} complex. a, Schematic of the haemagglutinin (HA) and Flag-tagged M4R-3C-A1R-3C-8×His construct. The most conserved residue for class A GPCRs (X.50 class A numbering[34]) are highlighted for each transmembrane domain in red. **b**, Purification step flowchart for the A_1R–G_{i2} complex. **c**, SDS–PAGE/western blot of samples obtained at various stages of A_1R–G_{i2} purification. A_1R and the G_{i2} heterotrimer were expressed separately in insect cell membranes. Addition of ADO initiated complex formation, which was solubilized by detergent. Solubilized A_1R and A_1R –G_i2 complex were immobilized on Flag antibody resin. Flag-eluted fractions were purified by size-exclusion chromatography (SEC). An anti-His antibody was used to detect Flag–A_1R-His and Gβ-His (red) and an anti-G_{i2} antibody was used to detect $G\alpha_{i2}$ (green). **d**, SDS–PAGE/Coomassie blue stain of the purified complex concentrated from the Superdex 200 Increase 10/30 column. **e**, Left, representative elution profile of Flag-purified complex on Superdex 200 Increase 10/30 SEC. Right, SEC fractions containing A_1R–G_{i2} complex (within dashed lines) were pooled, concentrated and analysed by re-running on Superdex 200 Increase 10/30 column. All images and SEC profiles are representative of more than three independent experiments.

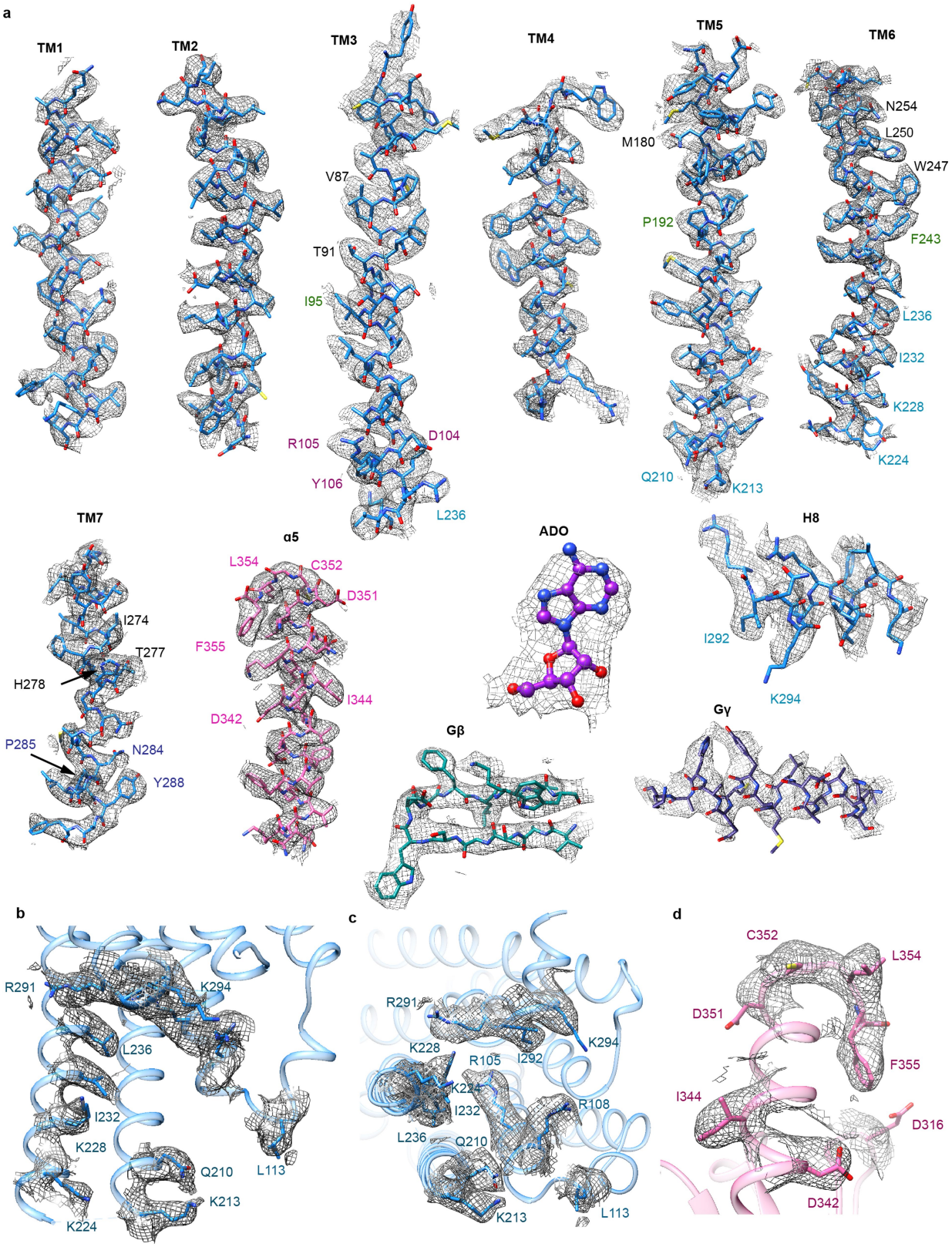


Extended Data Fig. 2 | See next page for caption.

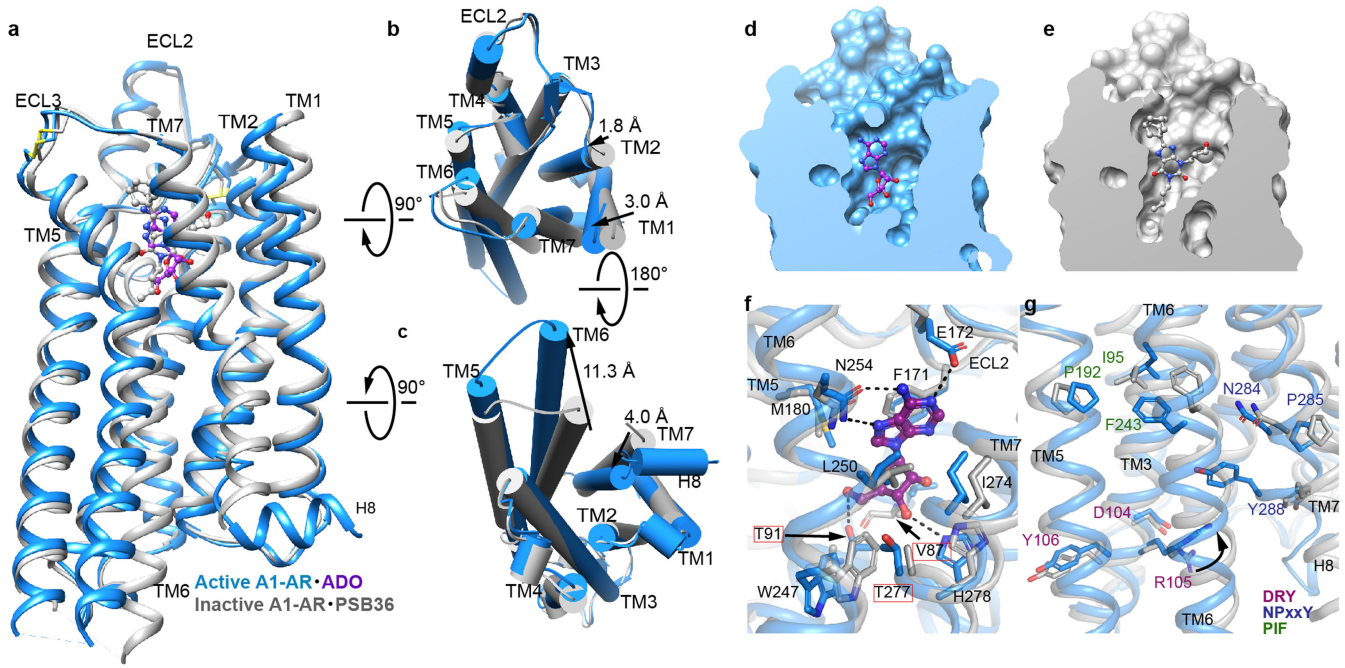
Extended Data Fig. 2 | Pharmacology of the A_1R construct and rationally chosen mutations. a, **b**, Competition between the antagonist [^{3}H]DPCPX and either unlabelled DPCPX (**a**) or ADO (**b**), in membranes expressing HA–Flag–3C-A1R-3C-8×His construct in the absence or presence of wild-type (WT) G_{i2} heterotrimer or dominant-negative (DN) G_{i2} heterotrimer. Data are normalized to [^{3}H]DPCPX binding in the absence of unlabelled competitor, with nonspecific binding determined in the presence of 1 μM of the antagonist, SLV320. **c**, ADO-mediated binding of [^{35}S]GTPγS as a measure of G-protein activation by the HA–Flag–3C-A1R-3C-8×His construct in High Five cells expressing receptor alone, or together with either wild-type or dominant-negative G_{i2} heterotrimer. **d**, **e**, [^{3}H]DPCPX competition assays (**d**) or inhibition of forskolin-stimulated cAMP accumulation (**e**), at the wild-type human A_1R or two key alanine substitution mutations stably expressed in CHOFlpIn cells. **f**, Changes in agonist (NECA) affinity (K_i) from the experiments shown in **d**. **g**, Changes in NECA signalling efficacy corrected for receptor expression (τ_c), determined from the experiments shown in **e**. Parameter estimates are the mean ± s.e.m. determined from 3 (**a**–**c**) or 6–48 (**d**–**g**) independent experiments performed in duplicate. $^{****}P < 0.0001$ (compared with wild type; one-way analysis of variance (ANOVA), Dunnett's post hoc test). Data for wild-type and N159A are replotted from Nguyen et al[28].



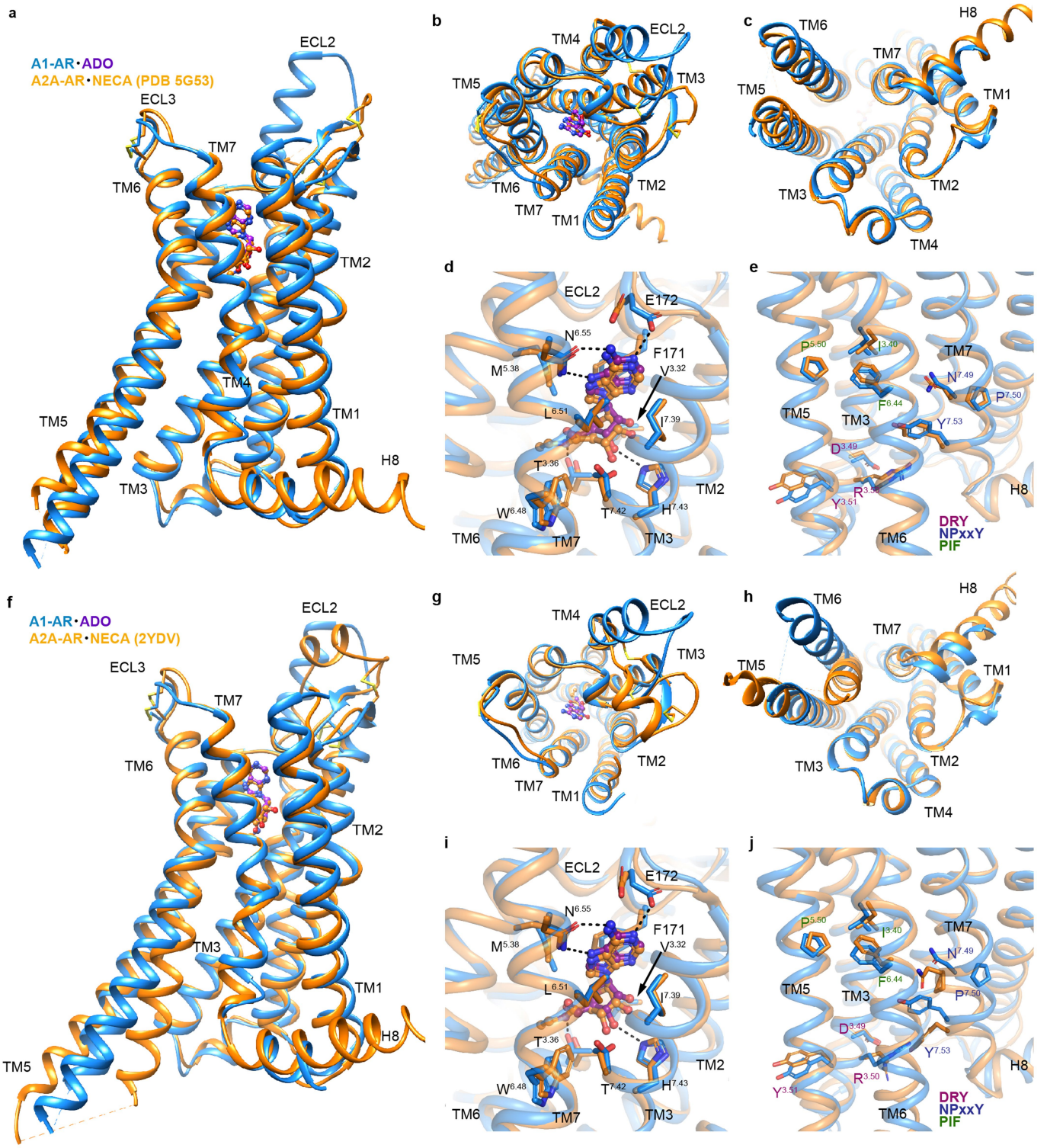
Extended Data Fig. 3 | Cryo-EM of the ADO–A_1R–G_{i2} complex. **a**, Representative VPP cryo-EM micrograph (of 3,220 recordings) of the ADO–A_1R–G_{i2} complex. **b**, Reference-free 2D class averages of the complex in LMNG and CHS detergent micelles. **c**, Gold-standard Fourier shell correlation (FSC) curves, showing the overall nominal resolution at 3.6 Å. **d**, FSC curves for the final model versus the final map and the half maps for overfitting validation (see Methods).



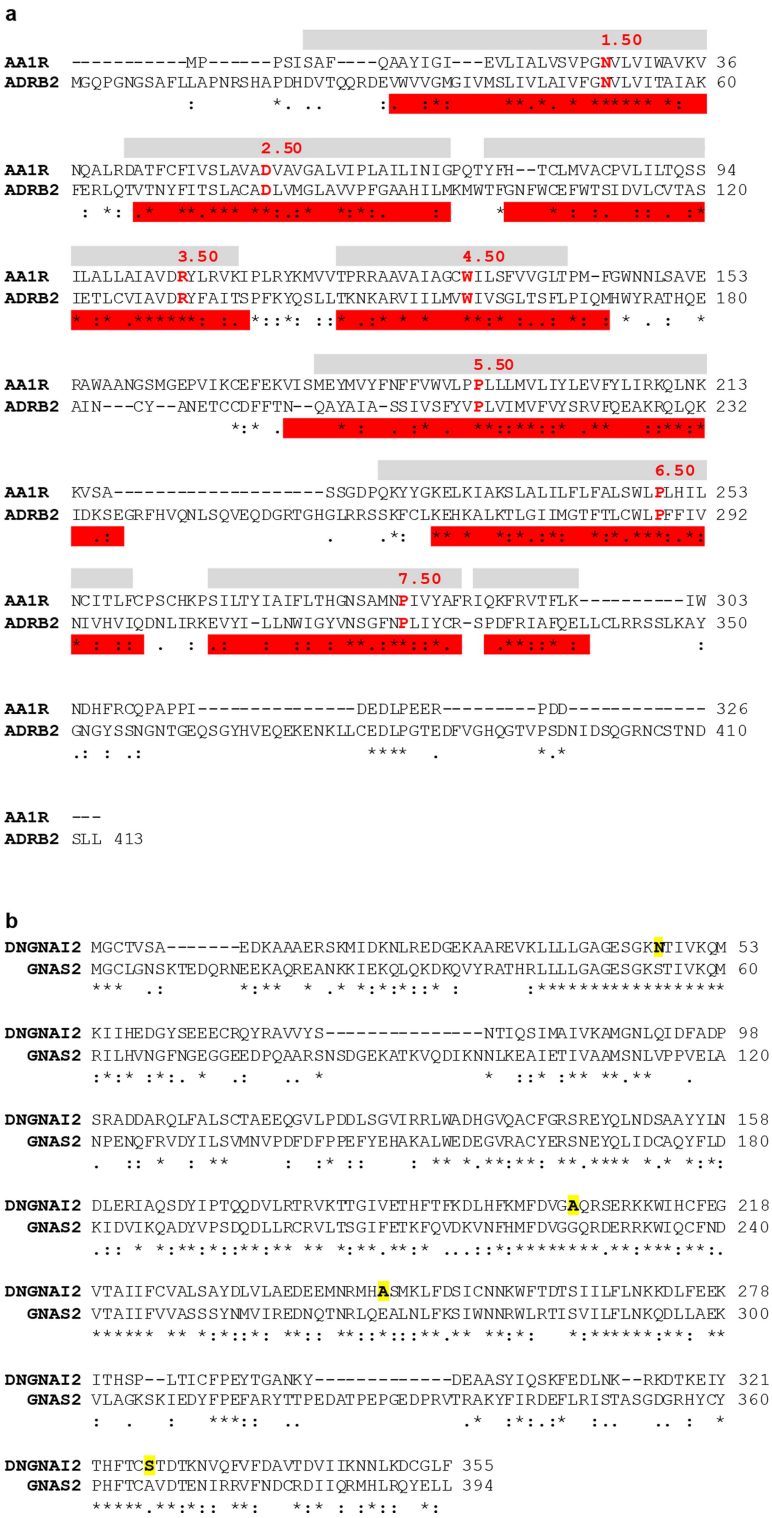
Extended Data Fig. 4 | Atomic resolution model of A_1R transmembrane domains, the Gα protein α5-helix, ADO, and representative regions of Gβ and Gγ in the cryo-EM density map. a, The molecular model is shown in stick representation and the cryo-EM map in mesh contoured at 0.06. **b**–**d**, A_1R residues (**b**, **c**,) and $G\alpha_{i2}$ α5-helix residues (**d**). The molecular model is shown in stick representation and the cryo-EM map in mesh contoured at 0.06.



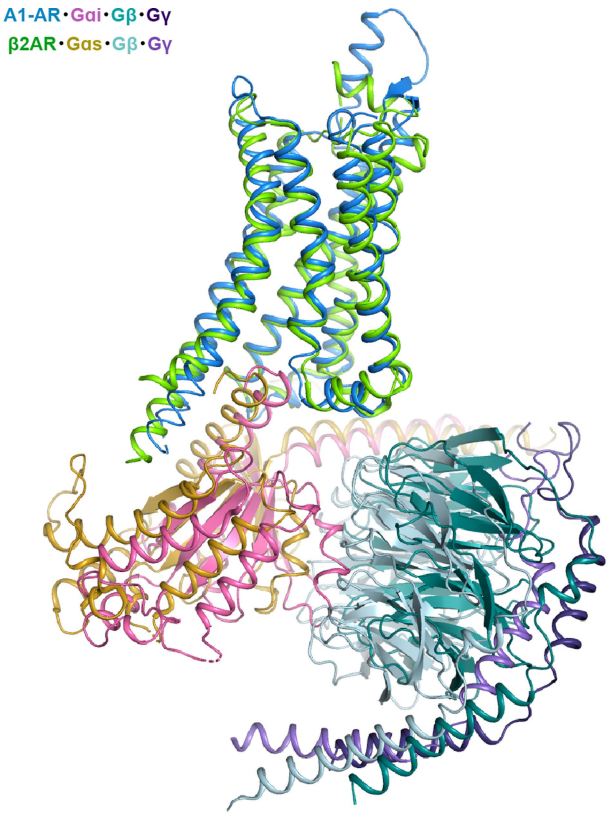
Extended Data Fig. 5 | Comparison of active and alternative inactive A_1R (PDB code 5NS2) structure. a–c, Side (**a**), extracellular (**b**) and cytoplasmic (**c**) view of the ADN–A_1R–G_{i2} structure (blue) compared to the inactive PSB36-bound A_1R (grey). **d**, **e**, Active ADO–A_1R (**d**) and inactive PSB36–A_1R (**e**) receptor surfaces sliced to show binding site cavity. **f**, Orthosteric binding site of the active A_1R–G_{i2} complex with ADO (purple ball and sticks). 'Toggle switch' $W247^{6.48}$ and residues within 4 Å of ADO are labelled and shown as sticks. Red rectangles highlight rotamer changes upon receptor activation. N, O and S atoms are coloured in blue, red and yellow, respectively. Dashed lines represent hydrogen bonds. **g**, GPCR motifs important for receptor activation (DRY motif, purple; NPXXY motif, blue; PIF motif, green).



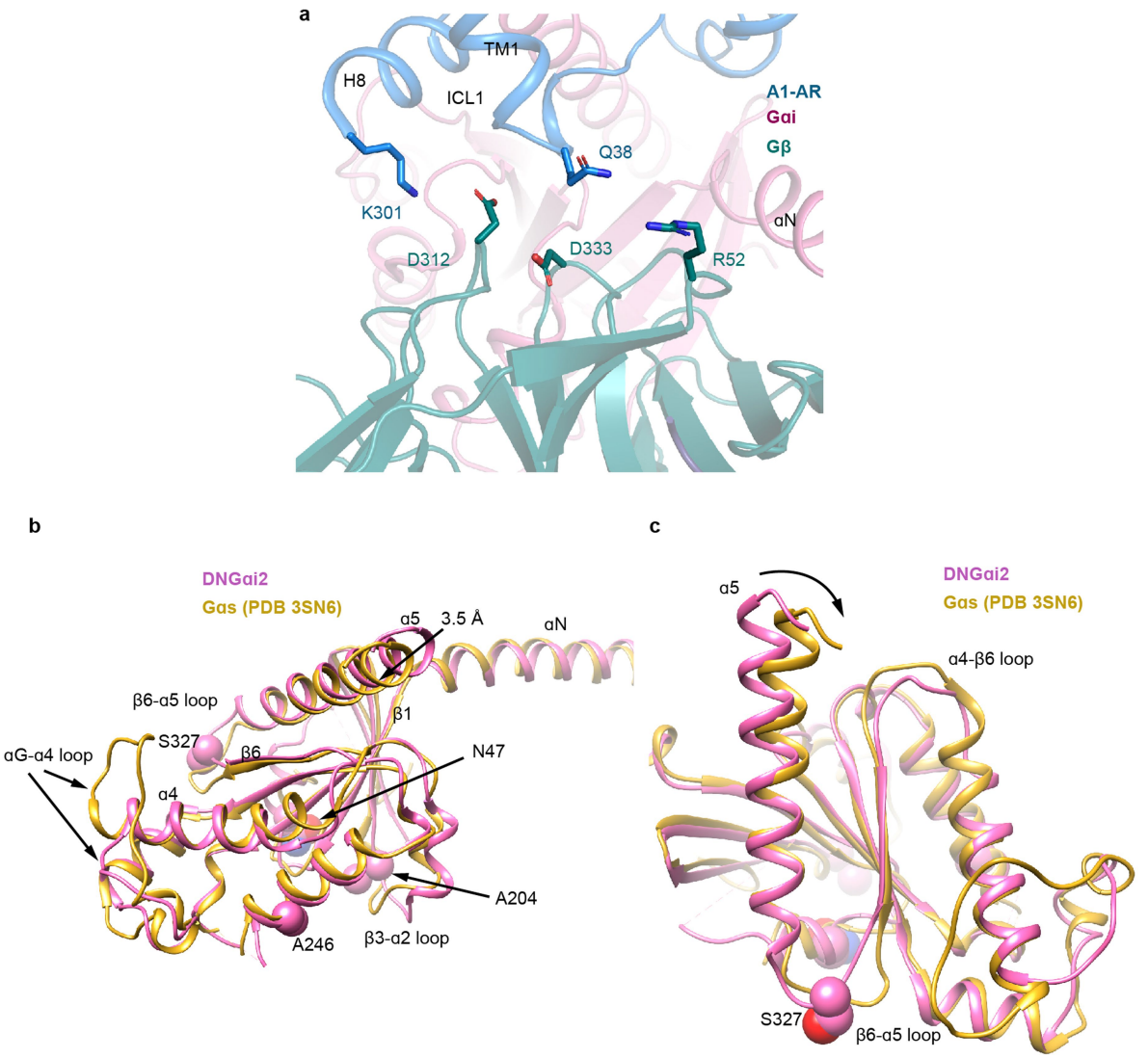
Extended Data Fig. 6 | Comparison of active A_1R with active $A_{2A}R$ (PDB code 5G53) or agonist-bound 'intermediate' state $A_{2A}R$ (PDB code 2YDV). a–j, Side views (**a**, **f**), extracellular views (**b**, **g**) and cytoplasmic views (**c**, **h**) of the active ADO–A_1R–G_{i2} structure (blue) compared to the active NECA–$A_{2A}R$–mini-G_s structure (**a**–**e**) or 'intermediate' NECA–$A_{2A}R$ structure (orange) (**f**–**j**). **d**, **i**, Orthosteric binding site of the active A_1R–G_{i2} complex with ADO (purple ball and sticks) or $A_{2A}R$ with NECA (orange ball and sticks). 'Toggle switch' residue $W^{6.48}$ and residues within 4 Å of ADO are labelled and shown as sticks. N, O and S atoms are coloured in blue, red and yellow, respectively. Dashed lines represent hydrogen bonds. **e**, **j**, Conserved class A GPCR motifs important for receptor activation (DRY motif, purple; NPXXY motif, blue; PIF motif, green).



Extended Data Fig. 7 | Alignments of A_1R with β_2AR or dominant-negative $G\alpha_{i2}$ with $G\alpha_s$. **a**, A_1R alignment with β_2AR. **b**, $G\alpha_s$ alignment with dominant-negative $G\alpha_{i2}$. Key Ballesteros–Weinstein numbers are shown in red. Grey bars indicate the positions of the α-helices in the A_1R–$G\alpha_{i2}$ structure, whereas red bars indicate these regions in the $\beta_{2A}R$–$G\alpha_s$ structure. Dominant-negative $G\alpha_{i2}$ point mutations are highlighted in yellow.



Extended Data Fig. 8 | Comparison of the A_1R–G_{i2} with β_2AR–G_s structures. Overlay of A_1R–G_{i2} with β_2AR–G_s (PDB code 3SN6) complexes. (A_1R–G_{i2} is coloured as in Fig. 1; β2AR is in green, $G\alpha_s$ is in gold, Gβ is in light cyan, Gγ is in light purple).



Extended Data Fig. 9 | View of key residues at the interface of A_1R and Gβ, and Gα conformations. a, A_1R is in blue and Gβ is in dark cyan. **b**, **c**, Different views comparing $DNG\alpha_{i2}$ and $G\alpha_s$ from A_1R and β_2AR (PDB code 3SN6) receptor-bound structures ($DNG\alpha_{i2}$, pink; $G\alpha_s$, gold). Spheres indicate the positions of the dominant-negative mutations on the $DNG\alpha_{i2}$ with N, O and C atoms coloured in blue, red and pink, respectively. The α5-helix bend and loops that are the most different between $G\alpha_{i2}$ and $G\alpha_s$ are indicated.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	A1-AR·G_{i2} (EMDB- EMD-7835) (PDB 6D9H)
Data collection and processing	
Magnification	47170
Voltage (kV)	300
Electron exposure (e–/Å^2)	50e/A^2
Defocus range (nm)	-500
Pixel size (Å)	1.06
Symmetry imposed	C1
Initial particle images (no.)	831,602
Final particle images (no.)	263,321
Map resolution (Å)	3.6
FSC threshold	0.143
Map resolution range (Å)	3.4 - 4.2
Refinement	
Initial model used (PDB code)	5UEN, 5G53, 3SN6, 5TDH
Model resolution (Å)	
FSC threshold	0.143
Model resolution range (Å)	3.6
Map sharpening *B* factor (Å^2)	-196
Model composition	
Protein residues	893
Ligand (ADO)	1
B factors (Å^2)	
Protein residues	21-125 (avr. 54)
Ligand (ADO)	44.5
R.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	1.2
Validation	
MolProbity score	1.42
Clashscore	4.18
Poor rotamers (%)	0
Ramachandran plot	
Favored (%)	96.6
Allowed (%)	3.4
Disallowed (%)	0

LETTER

https://doi.org/10.1038/s41586-018-0246-4

Macromolecular organic compounds from the depths of Enceladus

Frank Postberg[1,2,3,13]*, Nozair Khawaja[1,13], Bernd Abel[4], Gael Choblet[5], Christopher R. Glein[6], Murthy S. Gudipati[7], Bryana L. Henderson[7], Hsiang-Wen Hsu[8], Sascha Kempf[8], Fabian Klenner[1], Georg Moragas-Klostermeyer[9], Brian Magee[6,8], Lenz Nölle[1], Mark Perry[10], René Reviol[1], Jürgen Schmidt[11], Ralf Srama[9], Ferdinand Stolz[4,12], Gabriel Tobie[5], Mario Trieloff[1,2] & J. Hunter Waite[6]

Saturn's moon Enceladus harbours a global water ocean[1], which lies under an ice crust and above a rocky core[2]. Through warm cracks in the crust[3] a cryo-volcanic plume ejects ice grains and vapour into space[4–7] that contain materials originating from the ocean[8,9]. Hydrothermal activity is suspected to occur deep inside the porous core[10–12], powered by tidal dissipation[13]. So far, only simple organic compounds with molecular masses mostly below 50 atomic mass units have been observed in plume material[6,14,15]. Here we report observations of emitted ice grains containing concentrated and complex macromolecular organic material with molecular masses above 200 atomic mass units. The data constrain the macromolecular structure of organics detected in the ice grains and suggest the presence of a thin organic-rich film on top of the oceanic water table, where organic nucleation cores generated by the bursting of bubbles allow the probing of Enceladus' organic inventory in enhanced concentrations.

Two mass spectrometers onboard the Cassini spacecraft, the Cosmic Dust Analyzer (CDA) and the Ion and Neutral Mass Spectrometer (INMS), performed compositional in situ measurements of material emerging from the subsurface of Enceladus. These measurements were made inside both the plume and Saturn's E ring, which is formed by ice grains escaping Enceladus' gravity[16].

The CDA records time-of-flight (TOF) mass spectra of cations generated by high-velocity impacts of individual grains onto a rhodium target, with a mass resolution of $m/\Delta m \approx 20–50$[14,17]. Previous CDA measurements showed that about 25% of the ice grain spectra of the E ring, the so-called type-2 spectra, exhibit the presence of organic material[8,9,14]. A subgroup (about 3%) of type-2 spectra (see Methods sections 'Dataset' and 'Relative frequency of HMOC-type grains depends on impact speed and distance to Enceladus orbit'; Extended Data Table 1) is characterized by a sequence of repetitive peaks beyond 80 u (where u is the unified atomic mass unit), usually separated by mass intervals of 12 u–13 u (Fig. 1). In most cases, this sequence extends to the maximum mass nominally covered by the CDA, about 200 u. The broad and irregular shape of the peaks indicates that they are composed of multiple unresolved overlapping mass lines (Fig. 1a and Extended Data Fig. 1). The mass intervals between the peaks suggest organic species with an increasing number of carbon atoms (C_7 to C_{15}), which we refer to as high-mass organic cations (HMOCs). While an interval of 14 u would indicate the addition of a saturated CH_2 group to an organic 'backbone', the actual average mass difference of 12.5 u indicates the presence of predominately unsaturated carbon atoms.

In principle, each HMOC peak could be derived from a different parent molecule. However, a monotonic decrease without major intensity variations from 77 u to 191 u (Fig. 1a, Extended Data Fig. 1) rather indicate fragments from higher-mass parent molecules (outside the CDA's nominal mass range) and not a conglomerate of several molecules with masses below 200 u. This interpretation is consistent with the observation that the detection probability increases with impact speed (see Methods, 'Relative frequency of HMOC-type grains depends on impact speed and distance to Enceladus orbit'; Extended Data Table 1), increasing the available energy for fragmentation and ionization. A similar fragmentation pattern has been observed in impact experiments with polymers[18,19] containing aromatic subunits (Extended Data Fig. 2; see Methods, 'Inferring the origin of HMOC peaks in CDA spectra').

In addition to the nominal mass spectrum (1 u to ~200 u), the CDA records extended TOF spectra of the same individual grains up to about 8,000 u with resolution and sensitivity uniformly reduced by a factor of ten. These extended spectra exhibit abundant macromolecular cations with masses in excess of 200 u in most HMOC-type spectra (Extended Data Fig. 5). This further supports the interpretation that fragmentation of macromolecular parent molecules or networks creates the HMOCs.

The HMOC sequence always appears together with intense non-water signatures below 80 u (Fig. 1), offering further compositional and structural constraints. Although there is some variation (Extended Data Figs. 1, 4), non-water peaks are arranged in six groups at 15 u, 27–31 u, 38–45 u, 51–58 u, 63–67 u, and 77–79 u and, like the HMOC pattern, indicate a sequence of organic species with increasing number of carbon atoms. Organic mass lines below 45 u appear preferentially at odd masses (Fig. 1b), indicating a typical cationic hydrocarbon fragmentation fingerprint. However, from stoichiometry, the mass lines at 30 u, 31 u, 44 u and 45 u cannot be pure hydrocarbon cations and indicate O- or N-bearing cations from hydroxyl (CH_2OH^+, $CH_3–CH–OH^+$), ethoxy or carbonyl functional groups or nitrogen-bearing ions (for example, $CH_2NH_2^+$ and $CH_3–CH–NH_2^+$) (Fig. 1b, Extended Data Fig. 4).

A prominent peak preceding the HMOC sequence (Fig. 1b) indicates abundant cationic forms of a benzene ring, phenyl ($C_6H_5^+$, 77 u) and benzenium ($C_6H_7^+$, 79 u). The high abundance of these cations is highly diagnostic because the energetically favourable cationic aromatic structure would be tropylium cations[18–20] ($C_7H_7^+$, 91 u) (Fig. 1a). Formation of tropylium cations must thus be inhibited, as it requires aromatic precursor molecules without hydrogenated C-atoms (for example, alkyl groups) attached to the ring[20,21] (see Fig. 2, Extended Data Fig. 7 and Methods section 'Inferring the origin of HMOC peaks in CDA spectra').

A high abundance of aromatic structures is further supported by mass lines at 63 u–65 u ($C_5H_{3,4,5}^+$), 51 u–53 u ($C_4H_{3,4,5}^+$) and 39 u (C_3H_3) (Fig. 1b, Extended Data Fig. 1), which can be interpreted as coincident unsaturated benzene fragments[21] (Fig. 2, Extended Data Fig. 7).

[1]Institut für Geowissenschaften, Universität Heidelberg, Heidelberg, Germany. [2]Klaus-Tschira-Labor für Kosmochemie, Universität Heidelberg, Heidelberg, Germany. [3]Institut für Geologische Wissenschaften, Freie Universität Berlin, Berlin, Germany. [4]Leibniz-Institute für Oberflächenmodifizierung (IOM), Leipzig, Germany. [5]Laboratoire de Planétologie et Géodynamique, UMR-CNRS 6112, Université de Nantes, Nantes, France. [6]Space Science and Engineering Division, Southwest Research Institute, San Antonio, TX, USA. [7]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. [8]Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO, USA. [9]Institut für Raumfahrtsysteme, Universität Stuttgart, Stuttgart, Germany. [10]Applied Physics Laboratory, Johns Hopkins University, Laurel, MD, USA. [11]Astronomy Research Unit, University of Oulu, Oulu, Finland. [12]Wilhelm-Ostwald-Institut für Physikalische und Theoretische Chemie, Universität Leipzig, Leipzig, Germany. [13]These authors contributed equally: Frank Postberg, Nozair Khawaja. *e-mail: Frank.Postberg@geow.uni-heidelberg.de

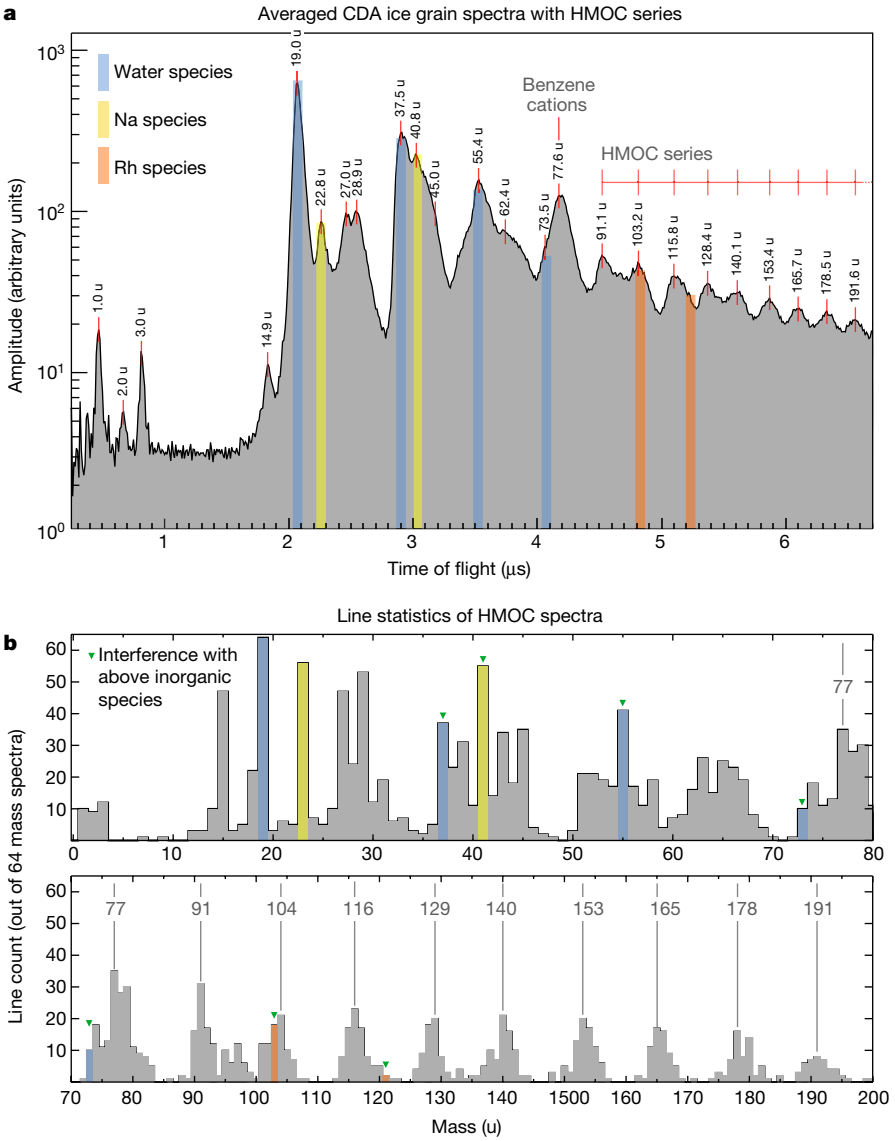


Fig. 1 | Co-added CDA HMOC spectrum and mass line histogram. **a**, Time-of-flight spectrum representing the average of 64 high-quality spectra (see Methods, 'Selection of 64 high-quality spectra for Fig. 1 and Extended Data Figs. 1 and 3'). The amplitudes of individual spectra were normalized and co-added. The corresponding masses of the peaks are labelled. The spectrum provides a representation of the average abundance of cation species. All signatures that are not colour-shaded are exclusively or mostly due to organic cations, as described in the text. Mass lines of exclusively inorganic origin are H_3O^+ at 19 u and Na^+ at 23 u. Mass lines where interference with inorganic cations is likely are: H_2O–H_3O^+ (37 u), H_2O–Na^+ (41 u), $(H_2O)_2$–H_3O^+ (55 u) and $(H_2O)_3$–H_3O^+ (73 u). The CDA's target material, rhodium, and its water cluster forms cations at 103 u and 121 u (Rh^+, Rh–H_2O^+), respectively, that interfere with the HMOC pattern, but only in fast impacts (see Extended Data Fig. 3 for a comparison of fast versus slow impacts and Extended Data Fig. 4 for examples of individual CDA spectra; for a semi-quantitative overview of spectral features in individual spectra in the dataset, see Extended Data Fig. 1b). **b**, Occurrences ('counts') of resolved mass lines and 'flank' peaks in 64 high-quality HMOC spectra. In contrast to the spectrum shown in **a**, the histogram makes no distinction of the peak amplitude and shows only the frequency of occurrence of resolved mass lines. Above 80 u the characteristic HMOC mass lines appear, preceded by the peaks from benzene-derived cations at 77 u and 79 u. A peak that can be identified at around 95 u in some spectra is in agreement with phenyl cations ($C_6H_5^+$, 77 u) forming a water cluster (C_6H_5–H_2O^+), preferentially at lower-impact speeds (Extended Data Figs. 1, 3). Organic mass lines below 45 u appear preferentially at odd masses, at 15 u, 27 u, 29 u, 39 u, 41 u and 43 u, indicating a preference for odd numbers of H atoms with the typical cationic hydrocarbon fragmentation fingerprint (CH_3^+, $C_2H_3^+$, $C_2H_5^+$, $C_3H_3^+$, $C_3H_5^+$ and $C_3H_7^+$). Although other interpretations for each individual line are possible (for example, HCN^+ (27 u), COH^+ (29 u) and CH_3CO^+ (43 u)), the overall pattern here suggests hydrocarbons or hydrocarbon fragments. In **b** all signatures with possible major contributions from inorganic species are colour-shaded as in **a** and marked by green triangles. In both panels, the absolute masses have an intrinsic uncertainty (absolute value) of ±1 u at 80 u and ±2 u at 180 u due to the limited calibration accuracy of the CDA in this high-mass regime. The mass intervals between peaks, however, are accurate to the integer level.

However, aliphatic fragmentation is also indicated by strong peaks at 27 u–29 u ($C_2H^+_{3,4,5}$) and 41 u–43 u ($C_3H^+_{5,6,7}$) and a less-pronounced peak at 15 u (CH_3^+). Mass lines at 55 u–57 u are in agreement with aliphatic C_4 species ($C_4H^+_{7,8,9}$), whereas aliphatic structures with more than four carbon atoms (for example, $C_5H^+_{9,10,11}$, with 69 u–71 u) are generally absent (Fig. 1).

Each spectrum also exhibits water-cluster cations of the form $H_3O(H_2O)_n^+$ (Fig. 1), typical for water ice impacts[14,17]. Evidently, an ice–organic mixture constitutes the bulk composition of these particles. Ion abundances are indicative of an organic fraction in ice grains up to the per cent level (Extended Data Fig. 7). Particle radii are mostly between 0.2 μm and 2 μm (Extended Data Table 1). Na^+ ions and

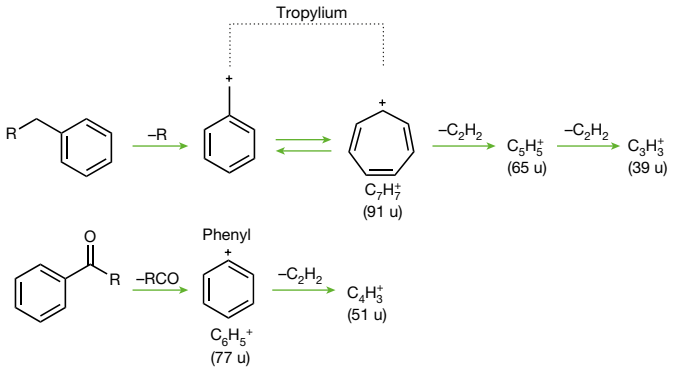


Fig. 2 | Formation of different aromatic cations. Although signatures in agreement with tropylium ions at ~91 u are present in every HMOC spectrum, they are on average about three times less abundant than the energetically favourable phenyl cation peak (Fig. 1a). If, upon breakup, the molecular structure allowed for an integration of a C–H group (top) into the highly stable aromatic resonance structure of the tropylium ion (91 u), then the formation of phenyl cations (77 u) would be suppressed (Extended Data Fig. 2). Only if the macromolecular structure does not allow this (bottom), phenyl (77 u) and benzenium (79 u, not shown) cations are the preferred single-ringed cations (also Extended Data Fig. 7). We note that fused benzene rings (such as PAHs) do not form either of these abundant single-ringed species (see Methods section 'Inferring the origin of HMOC peaks in CDA spectra' and Extended Data Fig. 8). Subsequent losses of neutral acetylene (C_2H_2) from phenyl and tropylium cations lead to smaller cationic fragments at 65 u, 39 u and 51 u[21], which are also seen in HMOC spectra (Fig. 1b).

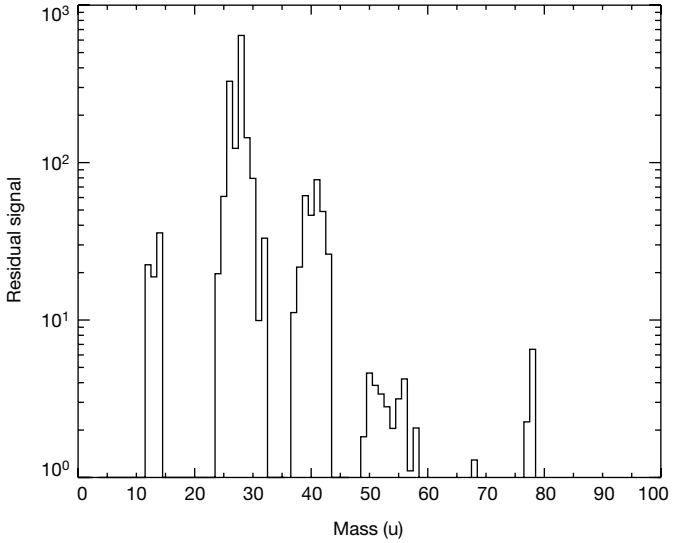


Fig. 3 | INMS organic fragmentation spectrum. By subtracting a plume-integrated INMS mass spectrum acquired at low velocities from a high-velocity spectrum, we obtain this residual signal (in arbitrary units), which we attribute to breakup products from high-mass species beyond the INMS mass limit at 99 u (Methods 'Residual spectrum from fast versus slow flybys'; Extended Data Fig. 11). Abundances of these high-speed fragments are normalized to the abundance of CH_4 (via its CH_3^+ ion at 15 u, which therefore has no signal in the residual spectrum). The species at 77 u and 78 u match the benzene-derived cations at 77 u and 79 u in the CDA spectrum. In contrast to the CDA, the INMS measures the composition of the neutrals and thus the neutral benzene ring (78 u) is expected to have the strongest signal in this spectrum. Many of the more abundant low-mass species are in agreement with aromatic fragmentation and oxygen-bearing parent molecules. Additional mass lines only seen at high-speed flybys (50 u–56 u) and much elevated signals in other masses (37 u–42 u and 25 u–28 u) indicate the unsaturated fragments C_4H_{2-8}, C_3H_{1-6} and C_2H_{1-4}, consistent with aromatic breakup products. Aliphatic fragmentation is required for the high-mass end of the C_2 and C_3 regions (29 u–30 u and 43 u). The overabundance at 28 u (CO), 30 u (probably H_2CO) and 31 u (probably CH_3OH) requires a contribution of oxygen-bearing fragments, consistent with CDA observations.

sodium–water clusters ($Na(H_2O)_n^+$) appear in most HMOC spectra, but at a much lower level than in spectra of type-3 grains, which are thought to be generated from frozen spray of Enceladus' salty ocean[8,9]. In HMOC-type spectra the Na-to-water ratio is similar to the low abundance found in type-1 particles thought to condense from salt-poor water vapour[8,22].

Cassini's INMS has measured the integrated composition of the Enceladean plume at several flyby speeds[12]. In contrast to the CDA, which records cations that form upon impact, the INMS simultaneously measures the composition of neutral gas entering the instrument aperture and volatile neutral molecules that are generated upon the impact of ice grains onto the instrument's antechamber[23] (see Methods, 'INMS data analysis'). There is a striking overabundance of organic species in spectra obtained at high flyby speeds (14–18 km s^{-1}) compared to those obtained at slower velocities (7–8 km s^{-1}) that we attribute to fragmentation of large organic parent molecules beyond the upper INMS mass limit of 99 u (Fig. 3). Given the low temperatures, species above 99 u are expected to be extremely depleted in the plume gas and originate from ice grains, probably of HMOC-type, entering the INMS aperture at high speed.

Mass lines at 77 u–78 u stand out, in full agreement with the CDA's inference of benzene species. Because these are only apparent at high flyby speed, they cannot stem from benzene itself but, like the other species in the residual spectrum, must be fragments from larger parent molecules. As in the case of the CDA, some of the smaller require oxygen-bearing fragments. Further INMS compositional analysis of signals extracted explicitly from ice grain impacts (see Extended Data Fig. 10 and Methods section 'INMS ice grain spectrum') provides further evidence that CO is the dominant fragment species at 28 u and that a N-bearing fragment (C_2H_3N) might be present.

Despite the relatively low mass range and resolution of the Cassini mass spectrometers, the measurements lead to the following key constraints. (1) The HMOC pattern in the CDA spectra and the extended spectra support the presence of organic molecules with masses clearly above 200 u. (2) The typical spacing of 12 u–13 u between HMOC peaks implies unsaturated cationic fragments with a ratio of C/H ≈ 2. Impact experiments with polystyrene[18,19] (Extended Data Fig. 2) indicate that the C/H ratio of the parent molecules might be lower (C/H ≈ 1) than in these fragments. (3) Prominent benzene-like species in the CDA and INMS spectra indicate the presence of abundant sub-structures of isolated benzene rings. From the INMS spectra, these features cannot originate from benzene itself, but are fragments of larger molecules. However, polycyclic aromatic hydrocarbon (PAH)-like fused rings do not form such fragments. Although not unique, the most parsimonious interpretation of these results is that these aromatic structures are parts of the same massive parent molecules that are responsible for HMOCs, which would explain their unsaturated nature. (4) From the suppressed formation of tropylium cations observed in the CDA spectra, we conclude that the rings are either connected to functional groups without carbon atoms or to dehydrogenated carbon atoms. (5) Unsaturated species at low masses are in close agreement with aromatic fragmentation. Aliphatic cations indicate saturated aliphatic structures with one to four C atoms arranged in parallel with the unsaturated (aromatic) structures. (6) Oxygen-bearing species in both the CDA and INMS spectra probably originate from hydroxyl, ethoxy or carbonyl functional groups. Nitrogen-bearing species are in good agreement with some features but do not provide a unique interpretation. Although no indication of other elements is observed in the organic structures, these cannot be ruled out.

Fragmentation of a single type of macromolecule may be responsible for all mass lines in the HMOC spectra. The parent substance would then be composed of cross-linked or polymerized aromatic and

aliphatic substructures with functional groups containing oxygen and probably nitrogen. Alternatively, contributions from lower-mass species might be mixed with macromolecules.

The Enceladean origin of these macromolecules is evident (see Methods sections 'Enceladus as the origin of HMOC parent molecules and exclusion of other potential sources' and 'Contamination of INMS spectra from previous measurements is unlikely'), and there is a multitude of speculative options for the genesis of these complex organics on Enceladus, some of which are outlined in Methods section 'Possible precursor scenarios for the observed complex organics'. However, the data do provide evidence of the formation of highly organic-enriched ice grains under quite specific conditions. Ice grains of very different salinity emerging from Enceladus have previously been identified in CDA mass spectra. Nearly pure water ice grains (type 1) can form from condensation of supersaturated vapour inside and above the ice vents[8,22,24]. In contrast, salt-rich ice grains (type 3) are thought to be frozen ocean spray[8,9], generated when bubbles of volatile gas (CO_2[25], CH_4 or H_2[12]) reach the water table and burst[26]. The HMOC-producing organic material (see Methods, 'Enceladus as the origin of HMOC parent molecules and exclusion of other potential sources') is detected in type-2 salt-poor ice grains, and thus cannot have formed directly from the salty ocean spray that preserves the composition of ocean water upon flash freezing[9]. Consequently, the observed organic compounds were not dissolved in ocean water when incorporated into ice grains. However, molecules with masses much larger than 200 u should not exist in the gas phase near the water table ($T \leq 0$ °C); hence, the organic substances cannot have condensed from the vapour either, and must have been solid when the grains formed.

The most plausible way to generate the observed grains is if the organic material exists as a separate phase, such as a thin film or layer of mostly refractory, insoluble organic species on top of (at least parts of) the oceanic water table located inside water-filled cracks in the ice crust (see Extended Data Fig. 12 and Methods section 'Deduction of an organic enriched layer at the Enceladean water table'). Bursting bubbles then disperse the organic film and, besides salty water droplets, produce droplets or flakes of organic material. When ascending in the icy vents, the organics become coated by water ice condensing from the vapour carrying the grains (Extended Data Fig. 12). Indeed, the limited tendency of benzene cations in HMOC spectra to cluster with neutral water molecules disfavours intimate mixing of organics with water and implies a core–shell grain structure. In a well mixed system, such a cluster at 95 u is much more pronounced (Extended Data Fig. 7).

Aerosol formation from bubble bursting is a well studied process in Earth's oceans, which are covered by an organic microlayer[26]. Organic-bearing sea spray serves as highly efficient nucleation cores of ice clouds over polar waters on Earth[27] and is found preferentially in the smallest aerosols, between 40 nm and 250 nm in size[26,28], whereas larger aerosols with sizes between 500 nm and 1,000 nm are either organics mixed with salt or do not contain organics[29]. Under Enceladus' low-gravity conditions, one would expect larger gas bubbles and therefore larger film drops. Spectra showing HMOCs are generated by ice grains with radii of around 1 μm (Extended Data Table 1). The presence of insoluble organic nucleation cores, a few hundred nanometres in size, would naturally explain the high organic content in HMOC grains. Larger, organic-free salty ocean droplets account for the salt-rich type-3 particles detected by the CDA. Bubbles ascend through tens of kilometres of ocean before reaching the surface. Like in Earth's oceans, organic substances can accumulate efficiently on the bubble walls[30], thus probing the oceanic organic inventory at depth (see Methods, 'Deduction of an organic enriched layer at the Enceladean water table').

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0246-4.

Received: 6 September 2017; Accepted: 23 March 2018;
Published online 27 June 2018.

1. Thomas, P. C. et al. Enceladus's measured physical libration requires a global subsurface ocean. *Icarus* **264**, 37–47 (2016).
2. less, L. et al. The gravity field and interior structure of Enceladus. *Science* **344**, 78–80 (2014).
3. Spencer, J. R. et al. Cassini encounters Enceladus: background and the discovery of a south polar hot spot. *Science* **311**, 1401–1405 (2006).
4. Spahn, F. et al. Cassini dust measurements at Enceladus and implications for the origin of the E ring. *Science* **311**, 1416–1418 (2006).
5. Porco, C. C. et al. Cassini observes the active south pole of Enceladus. *Science* **311**, 1393–1401 (2006).
6. Waite, J. H. Jr et al. Cassini ion and neutral mass spectrometer: Enceladus plume composition and structure. *Science* **311**, 1419–1422 (2006).
7. Hansen, C. J. et al. Enceladus water vapor plume. *Science* **311**, 1422–1425 (2006).
8. Postberg, F. et al. Sodium salts in E-ring ice grains from an ocean below the surface of Enceladus. *Nature* **459**, 1098–1101 (2009).
9. Postberg, F., Schmidt, J., Hillier, J., Kempf, S. & Srama, R. A salt-water reservoir as the source of a compositionally stratified plume on Enceladus. *Nature* **474**, 620–622 (2011).
10. Hsu, S. et al. Ongoing hydrothermal activities within Enceladus. *Nature* **519**, 207–210 (2015).
11. Sekine, Y. et al. High-temperature water-rock interactions and hydrothermal environments in the chondrite-like core of Enceladus. *Nat. Commun.* **6**, 8604 (2015).
12. Waite, J. H., Jr et al. Cassini finds molecular hydrogen in the Enceladus plume: evidence for hydrothermal processes. *Science* **356**, 155–159 (2017).
13. Choblet, G. et al. Powering prolonged hydrothermal activity inside Enceladus. *Nat. Astronomy* **1**, 841–847 (2017).
14. Postberg, F. et al. The E ring in the vicinity of Enceladus. II. Probing the moon's interior—the composition of E-ring particles. *Icarus* **193**, 438–454 (2008).
15. Waite, J. H. Jr et al. Liquid water on Enceladus from observations of ammonia and ^{40}Ar in the plume. *Nature* **460**, 487–490 (2009); corrigendum **460**, 1164 (2009).
16. Kempf, S., Beckmann, U. & Schmidt, J. How the Enceladus dust plume feeds Saturn's E ring. *Icarus* **206**, 446–457 (2010).
17. Postberg, F. et al. Discriminating contamination from particle components in spectra of Cassini's dust detector CDA. *Planet. Space Sci.* **57**, 1359–1374 (2009).
18. Goldsworthy, B. J. et al. Time of flight mass spectra of ions in plasmas produced by hypervelocity impacts of organic and mineralogical microparticles on a cosmic dust analyser. *Astron. Astrophys.* **409**, 1151–1167 (2003).
19. Srama, R. et al. Mass spectrometry of hyper-velocity impacts of organic micro grains. *Rapid Commun. Mass Spectrom.* **23**, 3895–3906 (2009).
20. Silverstein, R. M., Webster, F. X. & Kiemle, D. J. *Spectrometric Identification of Organic Compounds* 7th edn, 1–70 (John Wiley and Sons, Hoboken, 2005).
21. Dass, C. *Fundamentals of Contemporary Mass Spectrometry* 1st edn 210–238 (John Wiley and Sons, Hoboken, 2007).
22. Schmidt, J., Brilliantov, N., Spahn, F. & Kempf, S. Slow dust in Enceladus' plume from condensation and wall collisions in tiger stripe fractures. *Nature* **451**, 685–688 (2008).
23. Teolis, B. D., Perry, M. E., Magee, B. A., Westlake, J. & Waite, J. H. Detection and measurement of ice grains and gas distribution in the Enceladus plume by Cassini's ion neutral mass spectrometer. *J. Geophys. Res.* **115**, A09222 (2010).
24. Yeoh, S. K., Chapman, T. A., Goldstein, D. B., Varghese, P. & Trafton, L. M. On understanding the physics of the Enceladus south polar plume via numerical simulation. *Icarus* **253**, 205–222 (2015).
25. Matson, D. L., Castillo-Rogez, J. C., Davies, A. G. & Johnson, T. V. Enceladus: a hypothesis for bringing both heat and chemicals to the surface. *Icarus* **221**, 53–62 (2012).
26. de Leeuw, G. et al. Production flux of sea spray aerosol. *Rev. Geophys.* **49**, RG2001 (2011).
27. Wilson, T. W. et al. A marine biogenic source of atmospheric ice-nucleation particles. *Nature* **525**, 234–238 (2015).
28. Leck, C. & Bigg, E. K. Comparison of sources and nature of the tropical aerosol with the summer high arctic aerosol. *Tellus B* **60**, 118–126 (2008).
29. Gantt, B. & Meskhidze, N. The physical and chemical characteristics of marine primary organic aerosol: a review. *Atmos. Chem. Phys.* **13**, 3979–3996 (2013).
30. Porco, C. C., Dones, L. & Mitchell, C. Could it be snowing microbes on Enceladus? Assessing conditions in its plume and implications for future missions. *Astrobiology* **17**, 876–901 (2017).

Acknowledgements The research leading to these results received financial support from German Research Foundation (DFG) projects PO 1015/2-1, /3-1, /4-1 and ERC Consolidator Grant 724908—Habitat-OASIS (F.P., N.K, L.N., F.K. and R.R.), AB 63/9-1 (B.A. and F.S.), the Klaus Tschira Stiftung (M.T. and F.P.), NASA contract NAS703001TONMO71123, JPL subcontract 1405853 (J.H.W., C.R.G and B.M.), INMS science support grant NNX13AG63G (M.P.), NASA Habitable Worlds Program and JPL's RTD funding (M.S.G. and B.L.H.) and Academy of Finland project 298571 (J.S.).

Reviewer information *Nature* thanks J. Lunine and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions F.P. led the writing of the manuscript; N.K. and F.P. led the CDA data analysis with support from L.N.; G.M.-K. and R.S. designed the CDA observation; N.K., H-W.H., S.K., L.N. and F.P. did the programming and CDA data reduction; F.K., R.R., F.S., M.S.G. and B.L.H. conducted laboratory experiments; J.H.W. led the INMS observation; B.M. and M.P. conducted the INMS data reduction and analysis; F.P., J.S., C.R.G., M.T., G.T., G.C., M.S.G. and B.A. were responsible for the geophysical and geochemical interpretation of the data. All authors contributed to the discussion and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0246-4.

Correspondence and requests for materials should be addressed to F.P.

METHODS

CDA data analysis. *Short description of the CDA's chemical analyser subsystem*. The chemical analyser is the subsystem of the CDA[31] that provides chemical information about an impacting dust particle. Depending on its trajectory, the particle either hits the central rhodium target (chemical analyser target, CAT; diameter, 0.17 m), the surrounding gold target (diameter, 0.41 m) or the inner wall of the instrument. This work only deals with impacts on the CAT. If a dust particle impacts the CAT with sufficient energy, it becomes totally vaporized and partly ionized, forming an impact plasma consisting of target and particle ions, together with electrons and neutral molecules and atoms. The instrument separates this plasma, the positive component of which is linearly accelerated towards a multiplier about 19 cm away from the CAT, which is used to generate a TOF spectrum. The spectrometer is sensitive to positive ions only. The mass resolution, $m/\Delta m$, which is derived from laboratory experiments with the instrument, depends on the atomic masses of the ions. At 1 u, $m/\Delta m$ is 10, increasing to $m/\Delta m = 30$ at 100 u and up to $m/\Delta m = 50$ at 190 u—although these values vary strongly with impact conditions.

Acquisition of a spectrum in the high-rate sampling mode of the multiplier can be triggered by charge thresholds being exceeded on either the CAT or the multiplier. In the latter case, recording can be triggered up to several microseconds after the actual impact by the arrival of an abundant ion species; for water-dominated particles, these generally are hydrogen cations (H^+) or hydronium ions (H_3O^+). If triggered by the arrival of H_3O^+, mass lines of species with masses lower than 19 u, as well as the hydronium signature itself, do not appear in the spectrum. In this work, all spectra were triggered either directly by impact or by registering the H^+ signal at the multiplier.

The spectra are logarithmically amplified, digitized at 8-bit resolution and sampled at 100 MHz for a period of 6.4 μs after the trigger. The recording period of the high-rate sampling mode allows the detection of ions with masses of up to approximately 185 u, assuming that instrument recording is triggered by the impact itself, and about 215 u when recording is triggered by H^+. The acquisition of the extended spectrum starts after the high-sampling mode ends at 6.4 μs and is recorded with a sampling rate of 10 MHz up to a total time of 40 μs. The low number of data points corresponds to an extremely low mass resolution in the extended spectrum: one data point interval represents 7 u at 300 u and 20 u at 2,000 u. The extended spectrum frequently shows an instrument-artefact peak at 6.8 μs, which is ignored in our analysis.

Since the TOF is proportional to the square root of the mass-to-charge ratio of ions, in the ideal case its spectrum also represents a mass spectrum for identical ion charges. The ions created by impact ionization in the impact speed regime considered here (4–18 km s^{-1}) are almost exclusively singly charged. Unfortunately, the TOF is also influenced by the broad distribution of initial ion velocities, slightly varying flight paths and plasma shielding effects[32]. For that reason, species with identical masses are distributed over a range of sampling points and the mass resolution drops below integer values usually around 25 u–30 u. Therefore, we prefer to display original CDA spectra using a horizontal axis that shows the TOF, not the mass. The latter could incorrectly suggest an unrealistically high accuracy.

For a detailed description of the instrument, see Srama et al.[31]. The calibration routine is described in Postberg et al.[14,33]. For the dataset used in this work the calibration was done by using water and sodium mass lines, and a stretch factor of $a = 474$ ns was applied.

Dataset. All data used in this work are listed in Extended Data Table 1 and, like all CDA data, are archived on the Small Bodies Node of the Planetary Data System (PDS-SBN), at http://sbn.psi.edu/archive/cocda. We used 15 periods with good pointing of the CDA towards the E-ring dust ram direction between 2004 and 2008. This early period was chosen because there was negligible CDA contamination from salts deposited on the CDA impact target during deep Enceladus plume dives[17]. These 15 following periods (expressed in seconds), as recorded by the spacecraft clock, are:

1,477,548,025–1,477,746,968
1,489,004,979–1,489,118,561
1,498,469,760–1,498,505,448
1,506,223,870–1,506,436,625
1,509,311,996–1,509,357,627
1,511,686,278–1,512,480,582
1,514,090,477–1,514,186,815
1,515,560,669–1,516,328,727
1,519,522,033–1,519,612,908
1,521,596,023–1,522,361,262
1,543,798,083–1,543,831,204
1,544,838,694–1,544,916,311
1,557,517,652–1,557,520,425
1,589,065,736–1,589,083,623
1,605,559,719–1,605,566,052

Within these periods, a total of 7,353 spectra of impinging E-ring grains were investigated. A Lee filter was applied for better signal-to-noise ratio. Within the group of organic-bearing spectra (type 2[8,9,14]) we find a subgroup of spectra with a pattern of repetitive non-water peaks above 80 u (Fig. 1). We attribute this spectral pattern to HMOCs. The selection criteria are the following.

(1) At least five roughly well defined, equidistant peaks above 80 u with a signal-to-noise significance greater than 2σ.

(2) Water cluster peaks of the form $(H_2O)_nX^+$ (where X^+ is H_3O^+ or Rh^+, for example) are explicitly not considered for selection. Cases in which an HMOC pattern could be verified despite interference with water clusters were selected, but not considered for the sub-selection high-quality spectra (see Methods, 'Selection of 64 high-quality spectra for Fig. 1 and Extended Data Figs. 1 and 3').

(3) Isolated or irregular peaks above 80 u are not considered.

The much wider group of CDA type-2 grains[14] includes a multitude of organic-bearing ice grains observed in the plume and in the E ring. Other type-2 spectra not belonging to the type-2 subtype displaying HMOC patterns are outside the scope of this work.

Relative frequency of HMOC-type grains depends on impact speed and distance to Enceladus orbit. We find 83 spectra (1.1%) with apparent HMOCs, equivalent to about 3% of all type-2 spectra in the dataset (Extended Data Table 1). However, the probability of HMOC detection depends on the impact speed and thus the energy densities that the organic-bearing ice particles were exposed to upon impact. At high speeds (8–15 km s^{-1}) the overall frequency of HMOC detection is about 3% of all particles (~10% of type-2 spectra). At 6–8 km s^{-1} the HMOC detection frequency drops sharply to about 1% and goes down to 0.4% between 5 km s^{-1} and 6 km s^{-1} (see Extended Data Table 1). Below 5 km s^{-1} no HMOC was detected, equivalent to a probability of less than 0.1%. In the few spectra in the dataset recorded at speeds above 15 km s^{-1}, no HMOCs were detected (equivalent to a probability of <1.7%), so the speed range 8–15 km s^{-1} seems to be optimal for the CDA to record HMOC spectra. At speeds below 8 km s^{-1} the probability to produce fragments from the proposed parent molecules (outside the CDA mass range) decreases and becomes close to zero below 5 km s^{-1}. At speeds above 15 km s^{-1} the detection probability seems to decrease again, probably because at these high-impact speeds mostly small fragments with masses <80 u are produced, whereas the production of fragments that yield the typical HMOC pattern between about 80 u and 200 u (the defining feature of this CDA spectrum type) becomes less likely.

HMOC grains were detected in the plume (Extended Data Fig. 6) and in the E ring in a wide range of Saturnian distances from $3.8R_S$ to $14R_S$, where R_S is Saturn's radius, except in its outermost fringe, which extends[34] out to $20R_S$. However, the fraction of HMOC-type grains in the E ring increases closer to Enceladus' orbit. The chance to encounter an HMOC grain with intermediate impact speed of 6–8 km s^{-1} is about 2.5% at a Saturnian distance of $3.8R_S$–$5R_S$ and drops to 1.5% at $5R_S$–$7R_S$. Beyond $7R_S$, the detection probability decreases sharply to 0.5% and is 0.3% beyond $11R_S$. This clearly points at an Enceladus origin (see 'Enceladus as the origin of HMOC parent molecules and exclusion of other potential sources').

The effects of speed and distance are cumulative: in the optimal speed window and close to the orbit of Enceladus, the detection rate is highest, at about 4%. Thus, we can conclude that the actual fraction of ice grains carrying substantial amounts of complex organic material ejected into the E ring is at least 4% of the grains in the size range covered by the CDA (grain radius $r \approx 0.2$–2 μm). An instrument with an extended mass range (for example, up to 2,000 u) and higher resolving power than the CDA would probably detect complex organics with similar or higher probability at encounter speeds of ~5 km s^{-1} with the organic parent molecules being largely intact.

Selection of 64 high-quality spectra for Fig. 1 and Extended Data Figs. 1 and 3. For the peak statistics and co-added spectra in Fig. 1 and Extended Data Fig. 3 and the semi-quantitative spectra depicted in Extended Data Fig. 1, we used a subset of 64 spectra that had the highest quality. The 19 spectra that were not selected either had:

(i) an extended water-cluster pattern (for example, $(H_2O)_n$–H_3O^+ or $(H_2O)_n$–Rh^+), where strong peaks overlapped with the HMOC region above 100 u, or

(ii) an HMOC pattern with a low signal-to-noise ratio and poor peak definition that did not allow a reliable quantification of all features, or

(iii) an irregular baseline (instrument artefact) that did not allow a reliable quantification and would have spoiled the quality of the co-added spectra, or

(iv) an uncertain calibration, which did not allow an unambiguous assignment of peak masses.

In cases (i) and (ii), the organic fraction in the ice grains is probably lower than in other HMOC spectra.

In Extended Data Fig. 3 the dataset is split into 'fast' and 'slow' impacts (above and below ~10 km s^{-1}, respectively). The fast co-added spectrum is composed of 12 of the 64 high-quality HMOC spectra recorded after an impact with a velocity higher than ~10 km s^{-1}. In this case the speed was determined by the presence of

one or more hydrogen cations (H^+, H_2^+, H_3^+). The uncertainty of this method is clearly below the uncertainty achieved when using assumptions on the orbital elements, shown in Extended Data Table 1. The appearance of H^+ cations is a reliable indicator of impact speeds[17,35] in excess of about 12 km s^{-1}, and H_2^+ and H_3^+ cations appear already at slightly lower speeds (>10 km s^{-1})[17]. For 12 spectra in which one or more of these hydrogen mass lines were present, this reliable speed threshold indicator was used. The remaining high-quality spectra were co-added to produce the slow spectrum in Extended Data Fig. 3. Owing to their higher kinetic energy upon impact, low-mass cationic fragmentation species are more abundant there. Moreover, fast spectra were on average triggered earlier, mostly directly after impact (see 'Short description of CDA's chemical analyser subsystem'). Therefore, the high-rate recording interval of 6.4 μs stopped at a lower equivalent mass than for the lower-speed spectra (Extended Data Fig. 1), and the HMOC peak at ∼191 u was often not recorded for fast spectra.

Inferring the origin of HMOC peaks in CDA spectra. In the main text we argue that the HMOC pattern is indicative of highly unsaturated organic cations. A very similar pattern was also observed in CDA ground calibration experiments that used a dust accelerator[36] to produce impacts from micrometre-sized polystyrene beads (see figures 10 and 13 in Goldsworthy et al.[18]). The accelerator experiment was later repeated using an advanced TOF dust analyser with higher mass resolution[19] (Extended Data Fig. 2), giving identical results. A similar, but not identical, unsaturated cationic fragmentation pattern was observed in TOF-SIMS (secondary ion mass spectrometry) experiments with polystyrene[37]. The most plausible explanation for the fragmentation pattern and the ionic species observed is given in table 2 of ref. [37]. Although the spacing of 12.5 u in the HMOC pattern seems to suggest a C/H ratio of 2 or even higher, all three polystyrene experiments show that such an unsaturated cationic fragmentation pattern can also be achieved with parent molecules with a C/H ratio of only 1. These analogue experiments demonstrate that the C/H ratio of the cationic fragments could be different from that of the bulk material.

However, there is a noticeable difference between polystyrene spectra and HMOC spectra: in contrast to HMOC spectra, tropylium ions (91 u) form abundantly in impact ionization polystyrene spectra, whereas phenyl cations (77 u and 79 u) are relatively depleted. In Fig. 2 we demonstrate the reason: in polystyrene, polymer alkyl groups are attached to each of the aromatic monomers, which allow easy formation of tropylium ions at 91 u (Extended Data Fig. 2). This is also known from classical electron ionization mass spectrometry: abundant phenyl cations form only if the formation of tropylium ions is sterically difficult or impossible[20,21]. This leads to our conclusion that for HMOC parent species, single-ringed substructures are predominately attached to dehydrogenated carbon atoms or to functional groups without carbon.

In the dust accelerator experiments with polymers, an HMOC-like pattern was observed at impact speeds[18,19] of 4–6.5 km s^{-1}. Above this speed further fragmentation occurred. By contrast, CDA HMOC spectra are observed at impact speeds of about 5–15 km s^{-1}. The water ice matrix in which the HMOC parent molecules are embedded might explain this difference. It is well known from matrix-assisted laser desorption/ionization experiments that water protects large organic molecules from fragmentation very efficiently.

An alternative interpretation of the origin of HMOCs might be very large PAHs. However, in our laser-assisted analogue experiment with pyrene in a water matrix (see Extended Data Figs. 8, 9 and Methods section 'Laser dispersion analogue experiments for icy dust impacts'), we do not observe efficient fragmentation of the PAH. These findings generally agree with information from electron ionization (70 eV) mass spectrometry of a wide range of PAHs. By contrast, Stephan et al.[38] demonstrated a pattern similar to the one observed with the CDA above 100 u using TOF-SIMS of a mixture of high-mass PAHs. However, the applied energies were much higher (25 keV) and the PAHs were not embedded in a water matrix, which naturally protects organic molecules from breakup[39]. In any case, regardless of fragmentation at harsh ionization conditions, PAHs do not yield any monoaromatic peaks; hence, neither phenyl (77 u) nor tropylium (91 u) cations form[40]. Therefore, if PAHs have a role in the origin of HMOCs, other parent molecules must be responsible for the strong monoaromatic peak in the CDA and INMS spectra.

In principle, monoaromatic cations in the CDA spectra (77 u, 79 u) could originate from benzene mixed within the macromolecular structure responsible for the HMOCs. However, this is not possible for the benzene feature (78 u) seen in the INMS spectra because they exclusively appear at high-speed flybys and thus must be fragments of larger molecules. It is plausible that the monoaromatic features observed in both instruments originate from HMOC-type ice grains (see next section) and therefore also in CDA spectra we consider molecular fragmentation as the most likely source.

INMS data analysis. *Residual spectrum from fast versus slow flybys*. The INMS[41] measurement routinely integrates the composition of plume gas over the flyby to increase the signal-to-noise ratio. Using this methodology, volatile compounds released from ice grains that fragment and evaporate upon impact onto the instrument's antechamber[23] cannot be discriminated. In contrast to the CDA, which records cations, the INMS observes volatile neutral molecules that are in the vapour phase or are generated by ice grain impacts in the plume, and are subsequently ionized by the impact of 70-eV electrons in the ion source.

The residual spectrum presented in Fig. 3 represents the difference in composition and overabundance of organic material observed by the INMS during high-velocity flybys compared to low-velocity flybys through the Enceladus plume (Extended Data Fig. 11). Our interpretation of this spectrum is that low-speed flybys result in minimal fragmentation of the organic compounds within the ice grains. By contrast, high-speed impacts lead to fragmentation of large organic parent molecules beyond the INMS mass range of 99 u. Species with molecular masses above 99 u in the plume are expected to be extremely depleted in the gas phase and have to reside inside or on the ice grains. Based on the knowledge of the compositionally different ice grain populations that the CDA has collected over many years, HMOC-type grains entering the INMS's aperture are the most plausible source. The similarity in the composition of the fragments seen by the CDA and the INMS reinforce this assertion.

To obtain the spectrum shown in Fig. 3 we took the summed spectrum of the E5 flyby[15] and subtracted the average summed spectrum of the E14, E17 and E18 flybys. These represent the fastest (E5 at ∼18 km s^{-1}) and slowest (E14, E17 and E18 at ∼8 km s^{-1}) flybys through the plume for which the INMS has the best data. The three slow flybys have been combined into an average spectrum to compensate for their lower signal-to-noise ratio. Velocity-driven processes complicate the spectral comparison. H_2O–Ti reactions in the INMS antechamber increase with velocity owing to vaporization of Ti from the walls of the instrument's antechamber and create excess H_2 from H_2O, leaving TiO and TiO_2 as by-products[15]. Likewise, CO_2 fragments to become CO more readily as the velocity increases. Lastly, fragmentation products of high-mass organics probably enhance the plume gas signal for all masses dominated by organics in the INMS mass range. This makes it difficult to set an accurate reference point with which to compare the high- and low-velocity spectra. We have chosen to use the signal at 15 u in both the high- and low-velocity spectra as the common reference point because the CH_3^+ dissociative ion from CH_4 dominates the signal at this mass; furthermore, CH_4 is considered to be a plume-derived volatile species not considerably altered by fragmentation of high-mass organics. Choosing a mass of 15 u as the common reference point removes all the signals at this mass for the residual spectra, whereas in all other masses the signal is positive or negative. After matching the two spectra at their 15 u signal (Extended Data Fig. 11) and subtracting the low-velocity spectrum from the high-velocity spectrum, we set the amplitude scale of the residual relative to the noise floor of the low-velocity flybys (about 5 orders of magnitude lower than the H_2O signal at 18 u) so that only residual signal with a high signal-to-noise ratio is visible. The residual signal presented in Fig. 3 can be considered to be a conservative estimate of the high-mass organic fragmentation effect; if the CH_4 in the high-velocity flyby was increased from the low-velocity flybys, the residual signal would be even larger in amplitude.

Hydrocarbon species observed in this conservative estimate of the residual signal are discussed in the main text. Moreover, the residual spectrum suggests that at least 50% of the 28 u peak is from CO, given the upper limits placed upon N_2 and C_2H_4 in the fast E5 spectrum[15]. The oxygen-bearing species H_2CO and CH_3OH are suggested as the likely sources of the residual signal at 30 u and 31 u because they are the dominant species for these peaks in the E5 spectrum[15] and the residual retains more than 60% of the original E5 signal. This interpretation is strongly supported by the INMS ice grain spectrum.

INMS ice grain spectrum. To isolate spectral signals from icy grains when crossing the plume, we extracted those parts of the high-time-resolution, real-time INMS plume spectra (every 2.3 s) that registered the ice grains as spikes in the closed-source neutral (CSN) data[23,42]. The quadrupole mass analyser has a mass step integration period of 34 ms, whereas the vaporized gas from the impact dissipates in the antechamber in less than 3 ms. Therefore, the ice impacts appear as gas bursts or 'spikes' superimposed on the smoothly varying gas spectra that are mass-filtered by the mass step of the analyser. Thus, each spike can register only one unity-mass location per ice impact. By summing the spikes over three similar encounters (E14, E17 and E18) there are sufficient mass sampling statistics over the full sampled mass range to construct a pseudo-mass spectrum that represents the composition of the micrometre-sized grains.

Extended Data Fig. 10 shows this pseudo-mass spectrum of the grains measured by the INMS within 100 s of the time of closest approach. Each of the three encounters provided approximately 200 individual measurements or spikes to the composite spectrum. The vertical axis represents the relative count rate for each mass, because the counts for each mass are adjusted for the number of measurement opportunities (integration periods) for each mass. The scale of the plot is adjusted so that the minimum count rate is 1 count per integration period. The dispersion in the measurements obtained in the three encounters represents

the uncertainty on the count rates. The grain measurements are heavily concentrated near the closest approach, peaking with more than one spike per every ten integration periods at 2 u, which is the mass most sensitive to the grains. Analysis of the ice grain spectrum reveals a composition related to, but distinct from, the plume gas.

Two observations are particularly relevant to the high-mass organic material presented in this work (Extended Data Fig. 10). First is the clear dominance of CO at 28 u and the lack of N_2. By applying the well known quantities of dissociative species from the impact of 70-eV electrons in the INMS ionization chamber, CO_2 accounts for ~3% of the 28 u signal. N_2 has an upper limit of ~10% of the 28 u signal owing to the low signal at 14 u (N^+ dissociative ion) and may not be present at all. If present, C_2H_4 is limited to less than ~6% of the 28 u signal. This leaves a requirement for CO to be the dominant species at 28 u, which matches well the signal at 12 u and 14 u (C^+ and CO^{++} dissociative peaks). Although not a unique source, CO may be released from carbonyls in HMOC parent compounds by decarbonylation reactions during high-speed impacts of ice grains in the INMS. The caveat here is that the INMS integrates over all ice grains detected in the plume. Therefore, CO from carbonates or bicarbonates found in type-3 grains[9,11] might contribute here as well. The second observation is the likely presence of C_2H_3N. The signal pattern at 39 u–42 u suggests a species with its highest signal peak at 41 u and lower signal in the surrounding masses. C_2H_3N is the best candidate because it matches this 'stair step' pattern extremely well. C_2H_3N also appears to be present in the plume gas spectra, possibly because of contamination from a distribution of small grains indistinguishable from the incoming gas.

We note that in contrast to the spectrum in Fig. 3, these ice grain spectra were recorded at slow flyby speeds and show no signal with sufficient signal-to-noise ratio above 50 u. Unfortunately, ice grain spikes at the high-velocity flyby (E5) are not abundant enough to compose a similar spectrum, as shown in Extended Data Fig. 10.

Deduction of an organic enriched layer at the Enceladean water table. The observation of high concentrations of solid material in a water ice matrix that does not represent the ocean's salinity but is emitted from a saline ocean source, requires specific conditions during the formation of these grains. Because of the low temperatures near the water table where the ocean water is in contact with the ice crust, we conclude that the HMOC parent species are solid. The fact that the organics are embedded in a salt-poor water ice matrix indicates that they mostly are poorly soluble in water. In principle, primordial refractory organic substances trapped in the ice crust near the cracks, which become mobilized by the ascending vapour, could serve as an organic source. However, we disfavour this scenario because the crack walls are continuously coated by fresh ice condensing from the vapour flow[43,44].

We rather promote a scenario that is well known from ice cloud formation over polar waters on Earth[27]. There, organic aerosols of mostly biogenic origin[26] produced by bubble bursting serve as highly efficient nucleation seeds. When bubbles burst on Earth's oceans, an organic-free sea spray forms in parallel with pure organic aerosols and mixed phase organic-bearing sea spray[29,45]. The organic mass fraction of sea-spray aerosol has been consistently shown to be inversely related to aerosol size[29,46,47] . The purely organic end-members are found preferentially in the smallest aerosols[26,28] and are mostly water-insoluble[26,47]. As shown in the main text, the size of the organic nucleation cores in oceanic sea spray matches the CDA observation quite well.

Aerosol formation on Earth also provides a plausible analogue mechanism for the simultaneous production of salty ocean spray and organic aerosols: if the droplets are smaller than a few micrometres, then they can be thermally supported in water vapour with gas density slightly below the triple point against Enceladus' gravity. They do not fall back into the liquid, but are carried upwards through the ice vents by gas from evaporating water, following the pressure gradient into space. In parallel to the formation of smaller organic aerosols, larger salty ocean droplets form, which are later detected by the CDA as salt-rich type-3 particles in the plume and in the E ring. The analogue from Earth would also predict some larger mixed-phase particles that carry both ocean salts and complex organics, but these have not been identified by the CDA. One explanation could be that the macromolecular HMOC parent substance might be hydrophobic and would thus naturally avoid forming mixed-phase organic–sea-water aerosols.

On Earth, bubbles are generated mostly by breaking of waves[27]. On Enceladus, several volatile gases (CO_2, CH_4, NH_3 and H_2) have been detected in substantial concentrations in the plume[12] that will inevitably create bubbles when they rise though the ocean; these bubbles will then burst at the water table. In Earth's oceans, bubbles not only produce an aerosol, but are also very efficient in 'harvesting' organic molecules from the deeper oceanic environment by collecting these substances on their surfaces while ascending. The increase of relative organic concentrations observed near the surface of Earth's oceans by 2 to 3 orders of magnitude[48–50] suggests a selective transport of organic matter from the bulk seawater to the water table and then from the microlayer to atmospheric aerosols[27,29,51].

Possible precursor scenarios for the observed complex organics. There are many ways to explain the presence of complex organic materials in an icy moon in the outer Solar System. The two general categories of origin are accretion of primordial material and endogenic synthesis. In the former hypothesis, the organic carbon on Enceladus would predate the formation of the moon, and Enceladus would have acquired an organic inventory via its building blocks (icy planetesimals). The latter hypothesis would require hydrothermal systems inside Enceladus' porous rocky core[13] to produce complex organic molecules from small molecule precursors. This category can be broken down into several sub-categories characterized by molecular precursors. Organic compounds can be synthesized from more oxidized forms of carbon, such as CO_2, CO or formate[52,53]. More reduced simple organic species (formaldehyde, methanol and HCN) can also serve as feedstock for the synthesis of more complex organic compounds[54,55]. Methane is a relatively inert species, so it may be a less favourable carbon source, unless prolonged metamorphism or radiation chemistry is involved. Both abiotic and biotic processing of these precursors is possible. A mixture of spatially distinct (for example, ocean versus rocky core) sources is also not to be excluded at the present state of knowledge.

As an example of endogenic synthesis, relatively oxidizing hydrothermal conditions may promote the conversion of simple primordial organics into reactive unsaturated compounds, such as quinones, (poly)phenols or aldehydes, which may in turn polymerize (possibly in the presence of catalytic minerals) to form relatively hydrogen-poor macromolecules. Macromolecules on Enceladus that contain aromatic units with connecting short aliphatic chains that include more or less oxidized functional groups may resemble some humic substances on Earth. On Earth, several pathways exist for the formation of humic substances during the decay of biogenic complex organic matter[56]. However, these macromolecules may also be formed via synthetic routes, such as radical polymerization of phenolic compounds in the laboratory. Hänninen et al.[57] showed that humic acid-like polymers can be synthesized from homogeneous, well defined starting materials under oxidative conditions. The polymers display clear signatures of phenolic, aromatic (olefinic) and carboxyl carbons, and carboxyl carbons are present regardless of whether the monomeric unit possesses a free carboxyl group or not. On this basis, it was concluded that a partial de-aromatization occurs during the oxidative polymerization of o- and p-diphenolic compounds. Hänninen et al.[57] also reported that the ^{13}C nuclear magnetic resonance spectra and other features of the synthesized polymers resemble to a large extent the spectra of humic acids. The characteristics of these macromolecules are consistent with our observations from the HMOC spectra.

One can also consider a primordial origin of organic materials, which would be the simplest possibility. If the rocky materials that were accreted by Enceladus were analogous to CI-, CM- or CR-group chondrites or refractory cometary solids, then a substantial organic inventory is inescapable. CI chondrites in particular contain ~2 wt% insoluble organic carbon[58]. This insoluble organic matter (IOM)[59–63] is considered to be partly of primordial origin (that is, inherited from the interstellar medium) and partly modified by early hydrothermal processing on carbonaceous chondrite parent bodies. These asteroidal small bodies or planetesimals accreted within a few million years after the Solar System formation, heated up as a result of radioactive decay (mainly from ^{26}Al) and persisted at elevated temperatures, causing hydrothermal alteration and metamorphism for at least a few tens of millions of years after Solar System formation. While analytical data for comets are much more limited, recent results from comet 67P indicate a much higher organic carbon content of ~30 wt% in dust particles[64,65]. To put these extraterrestrial numbers into perspective, we note that hydrocarbon source rocks on Earth have an average total organic carbon value of ~2 wt%[66].

Enceladus organic matter may be similar to chondritic IOM. The latter shows some compositional variability: particularly in classes like CV and CO, which experienced strong thermal metamorphism at temperatures of about 400 °C–600 °C, both the H/C and O/C ratios are lower (down to 0.1 and 0.05, respectively), whereas in more primitive classes (CI, CM, CR) mild aqueous activity (100 °C–200 °C) prevailed and allowed the preservation of O/C ratios as high as 0.23 and H/C ratios[59] of 0.8. IOM in the CI, CM and CR classes is considered to be the most primitive type of IOM and has an average composition of $C_{100}H_{70}O_{22}N_3S_7$. It is macromolecular, and most of the carbon is incorporated in small aromatic structures, with about 20%–30% of the carbon constituting aliphatic bonds. The maximum length is 7 carbon atoms for aliphatic bridges between aromatic units and 4 carbon atoms for side chains with a free end[60,61]. Functional groups containing O, N and S are also abundant (see Fig. 2 in Remusat[61]). These characteristics are consistent with our observations from the HMOC spectra.

Because this subtype of IOM is thought to be the product of mild thermal and aqueous alteration, on the basis of present information it is impossible to say if this alteration already happened inside icy planetesimals under the influence of ^{26}Al heating or only later on Enceladus. There, under the influence of tens of millions of years of hydrothermal processing[13], more primitive primordial organics could have been transformed into a different organic mixture.

In the following, we will elaborate further on scenarios that invoke some level of processing of primitive organics on Enceladus. At a depth of about 60 km below the surface, the ocean comes in contact with Enceladus' core. The core's very low density (~2,500 kg m^{-3}) suggests ~20% porosity[2,13]. Owing to the low pressures and modest temperatures in Enceladus' interior, an unconsolidated core with such porosities can be inherited from the accretion and differentiation process. Tidal forces help to maintain this fragmented state so that ocean water can percolate through it, up to the present day. In Enceladus' interior, temperatures of 100 °C or more (increasing towards the core centre) can be maintained over at least tens of millions of years. Depending on the permeability of the core, the entire ocean could be processed at temperatures higher than approximately 100 °C within[13] 25–250 Myr.

While primordial oxygenated and nitrogenous species, such as alcohols, carboxylic acids, amines and nitriles, may be more soluble in liquid water and can be leached into the subsurface ocean of Enceladus, larger polymeric organics would exist as a separate organic layer. If not heated above ~300 °C, these organics might remain largely unaltered over geological timescales. Intense interaction with water, however, might have oxygenated the macromolecules on one end to form micellar structures (hydrophilic exterior and hydrophobic interior) and preserved their structural integrity over the lifetime of Enceladus.

Alternatively, a large accreted inventory of IOM-like material could establish the basis of a more evolved organic factory inside Enceladus' rocky core, similar to oil- and gas-generating sedimentary basins on Earth. To make this factory operational and enable consistency with the organic observations in the plume, organic compounds must be mobilized from source rocks. Hydrothermal activity[10,12] could facilitate this process. Cooking organic matter (metagenesis) in the core would crack it into smaller constituents, which may then be transported by fluid flow[13] as dissolved species, liquid droplets or entrained particulates (of particular relevance to the observations). A possible Earth analogue for this general scenario is Guaymas Basin in the Gulf of California[67]. This type of scenario has also been suggested for the interior of Pluto[68].

Is a scenario of processing primordial organic materials in Enceladus' core consistent with the chemical and structural features of compounds detected in the plume? Before proceeding, it is important to emphasize that it is still an open question whether we are seeing features that are representative of bulk organic materials in the subsurface. Processes that could fractionate organic compounds from their hypothesized source region to our instruments include expulsion from the core, (bio)degradation, hydrophobic phase separation in the ocean, plume outgassing and impacts during high-speed flybys. In light of these considerations, it seems prudent to focus on broad chemical characteristics.

Our data show the presence of unsaturated carbon and in particular the benzene ring. The former finding agrees qualitatively with elevated temperatures that promote entropically driven (more product than reactant molecules) dehydrogenation reactions. Countering this effect might be high concentrations of H_2 in hydrothermal fluids[12]; however, hydropyrolysis experiments demonstrate that aromatic carbon can coexist with abundant H_2[69]. The outcome on Enceladus presumably depends on the conditions of temperature and H_2 concentration, as well as the duration of heating and the availability of hydrogenation catalysts such as nickel. If primitive organic matter in chondrites[58] and comets[70] can serve as a guide, it can be expected that Enceladus would have accreted organic materials containing benzene rings. From the perspective of a primordial origin, it therefore makes sense to find the benzene ring as part of the organic structures at Enceladus. It is unknown how much thermal maturation has occurred, but complete graphitization can be ruled out because the presence of organic-bound hydrogen is implied by the data. This imposes a limit on the thermal history of the organic source rocks, which seems consistent with the persistence of a low-density core rich in hydrated silicates[2]. These organics have not been overcooked.

Two other compositional features are relevant to a general discussion. First, the likely presence of oxygen- and nitrogen-bearing functional groups is consistent with inheritance of these heteroatoms in accreted organic matter (O/C ≈ 0.2, N/C ≈ 0.04)[58,70] or their subsequent incorporation into organic structures by hydration or amination reactions (for example, of C=C or C≡C) on Enceladus. The latter processes would be facilitated by the availability of liquid water and ammonia in the interior[12]. Having heteroatoms also constrains the degree of thermal maturation. The second feature that we wish to highlight is the presence of methane in the plume[12]. It must be noted that there are multiple ways to explain the origin of methane[71], but it is appealing to envision a common mechanism for the formation of the full spectrum of organic compounds found at Enceladus. The analogy to petroleum geochemistry on Earth implies that thermal processing of organic materials will inevitably produce some CH_4 accompanying more complex organics[66]. The exact quantity of thermal gas will depend on the nature of the organic source material and the environmental (for example, redox) conditions. At Guaymas Basin, hydrothermal fluids have very high concentrations of CH_4 (~60 mmol kg^{-1})[72]. If the rocky core of Enceladus is also organic-rich and heated sufficiently, then this becomes a plausible scenario.

Enceladus as the origin of HMOC parent molecules and exclusion of other potential sources. *Photolysis in the E ring.* Most individual ice grains reside in the E ring for months to a few decades before they collide (either with a moon or the main rings) or before they are completely eroded by plasma sputtering. The oldest populations are generally located further away from their origin, Enceladus, in the outer E ring[73,74]. In principle, one has to consider that either HMOC parent molecules are not from Enceladus but evolved by photolysis from originally simple organics to the observed complex compounds in the E ring, or that the complex organics are from Enceladus but have been severely altered—for example, by dehydrogenation. However, several observations are not in agreement with these scenarios.

- The CDA observed a larger proportion of HMOC-type ice grains in the (young) inner E ring than in the (old) outer E ring (see Methods, 'Relative frequency of HMOC-type grains depends on impact speed and distance to Enceladus orbit'). This indicates degradation of the observed complex organics with time in the E ring, rather than their generation.
- During Cassini's E17 flyby, one freshly ejected HMOC-type ice grain was observed inside the plume (this was not included in the compositional analysis of this study, but it is shown in Extended Data Fig. 6). Furthermore, the CDA observed many HMOCs at locations close to Enceladus' orbit, where most grains had been ejected from the plume only a few days to few months ago.
- The INMS observations were directly made in the plume with freshly ejected material.

Instrument contamination. Intrinsic instrument contamination can be excluded as a source for the abundant organics in the spectra[17]. Of concern are the large PAHs in Titan's atmosphere, which may have deposited onto the surface of the CDA's impact target during close Titan flybys. Several pieces of evidence point against cross-contamination from Titan:

- The first HMOC spectra were already detected in 2004 and early 2005, before Cassini's first close flyby of Titan.
- The CDA has a decontamination device that was used between Titan flybys, which heats the CDA impact target to 100 °C for several hours.
- No buildup of organics can be observed in spectra of E-ring grains over time.
- The fact that only about 1% of impacts in the E ring show such a massive organic signature is not in agreement with a coating of organic contamination on the impact target.
- The trend to observe HMOC-type grains at a higher frequency closer to Enceladus' orbit than in the outer E ring (see 'Relative frequency of HMOC-type grains depends on impact speed and distance to Enceladus orbit') is not in agreement with a contamination origin.
- Thousands of spectra from impacts of dust populations outside the E ring (stream particles, exogenous dust and main-ring dust) never show HMOC features.

Contamination of INMS spectra from previous measurements is unlikely. The multiple environments encountered by the INMS require an evaluation of the potential for one environment to deposit refractory material onto the surfaces of the CSN antechamber and for that material to be released and measured at a later encounter. Of particular concern are the large PAHs in Titan's atmosphere and Enceladus' ubiquitous nanograins, which may fragment material deposited onto the antechamber surfaces, sputter the organic products, and thus release volatiles into the INMS.

The strongest evidence that high-speed nanoparticles do not create false signals in the INMS spectra is that the amounts of PAHs encountered at Titan are insufficient to produce the INMS measurements obtained during encounters such as E5. At Titan, both the INMS and the CAPS measure PAHs or their fragments, with benzene observed by both instruments. Waite et al.[75] and Lavvas et al.[76] analysed these measurements to estimate the abundance of PAHs as functions of altitude and of mass. Combining these estimates with the conservative assumption that every PAH molecule that enters the INMS CSN aperture during the Titan encounters is then released during E5, the resulting signal would be more than an order of magnitude below the level measured by the INMS at the fast Enceladus flyby (E5), as detailed in the following paragraph.

The benzene density in Titan's atmosphere is well modelled by an exponential with a scale height of 15–20 km and a density of approximately[75] 10^5 molecules per cubic centimetre at 950 km. Before E5, there were thirteen Titan encounters with the INMS pointed to accept atmospheric neutrals, and approximately 10^{11} higher-mass molecules entered the INMS CSN antechamber during those encounters. To produce the E5 measurements, the CSN antechamber maintained a density of less than 10^7 molecules per cubic centimetre for at least 40 s for the heavier species. With a residence time of only a few milliseconds, more than 10^{12} heavy molecules were required in the CSN antechamber during E5. The number of molecules encountered at Titan is thus at least a factor of ten below the number of molecules required in the CSN antechamber to produce the higher-mass count rates observed during E5.

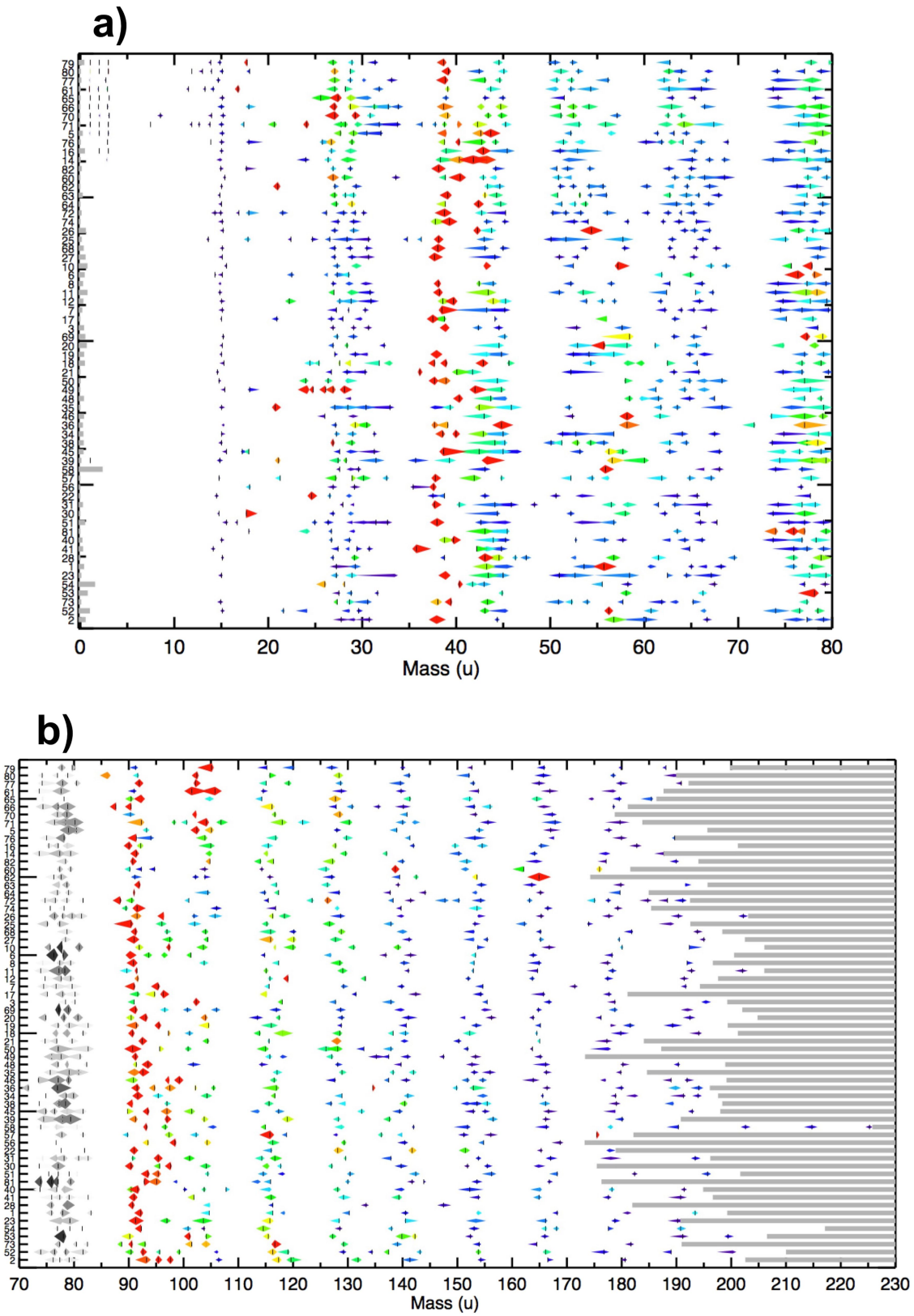
There is further evidence that the INMS spectra are unaffected by previous encounters: the CSN spectra at Enceladus do not change over many years, despite scores of Titan encounters during that period. Furthermore, there is no apparent dependency of Enceladus' spectrum on whether a Titan encounter occurred between sequential Enceladus flybys.

Laser dispersion analogue experiments for icy dust impacts. Extended Data Fig. 9 shows a sketch of the experimental setup in Heidelberg. A liquid water beam, in which the tested substances are dissolved, is injected into a high vacuum (5×10^{-5} mbar) through a quartz nozzle (with opening radius 6–8 μm). The quartz nozzle is mounted on a three-axis manipulator. A high-performance liquid-chromatography pump (model 300c, Gynkotek) is used to keep the flow rate constant at 0.17 ml min^{-1}. The liquid beam flow is stable for ~3 mm downwards and then disintegrates into droplets. To maintain the high vacuum, the liquid pours into two liquid-nitrogen-cooled cryotraps. To perform the measurements, the water beam is hit by a pulsed infrared laser (Opolette HE 2731, OPOTEK; 20 Hz, 7 ns pulse length), which operates at a wavelength of approximately 2,850 nm and a pulse energy of up to 4 mJ. The laser is directed and focused onto the liquid by three CaF_2 lenses and a gold mirror. The wavelength is chosen to specifically excite the OH stretch of the water molecules. When the laser energy is absorbed by the water beam, the water is rapidly heated and disperses explosively into fragments. In this way cations, anions, electrons and neutral molecules are created[39,77].

Depending on the operation mode, cations or anions that pass through a skimmer are analysed in a reflectron TOF mass spectrometer (Stefan Kaesdorf). This mass spectrometer operates according to the principle of delayed extraction. The delay time is adjusted with the aid of a pulse generator (model DG535, Stanford Research Systems). By adjusting the delay time, ions with a distinct initial velocity are selected for detection by a microchannel plate detector. In combination with different delay times, variable laser intensities are used to simulate different impact velocities of ice grains onto the CDA's target (F.K. et al., manuscript in preparation). The detected signals are intensified by a preamplifier, visualized by a 12-bit digitizer card (Acqiris) and recorded using LabVIEW. Flow injection of the tested solutions is accomplished with an injection valve (model MX9925, Rheodyne). 500 single spectra are averaged to achieve the co-added spectra, as presented in Extended Data Figs. 7, 8. The chemicals (pro analysi) were purchased from Sigma Aldrich. All solutions were freshly prepared with doubly distilled H_2O in lockable 50-ml sample cups.

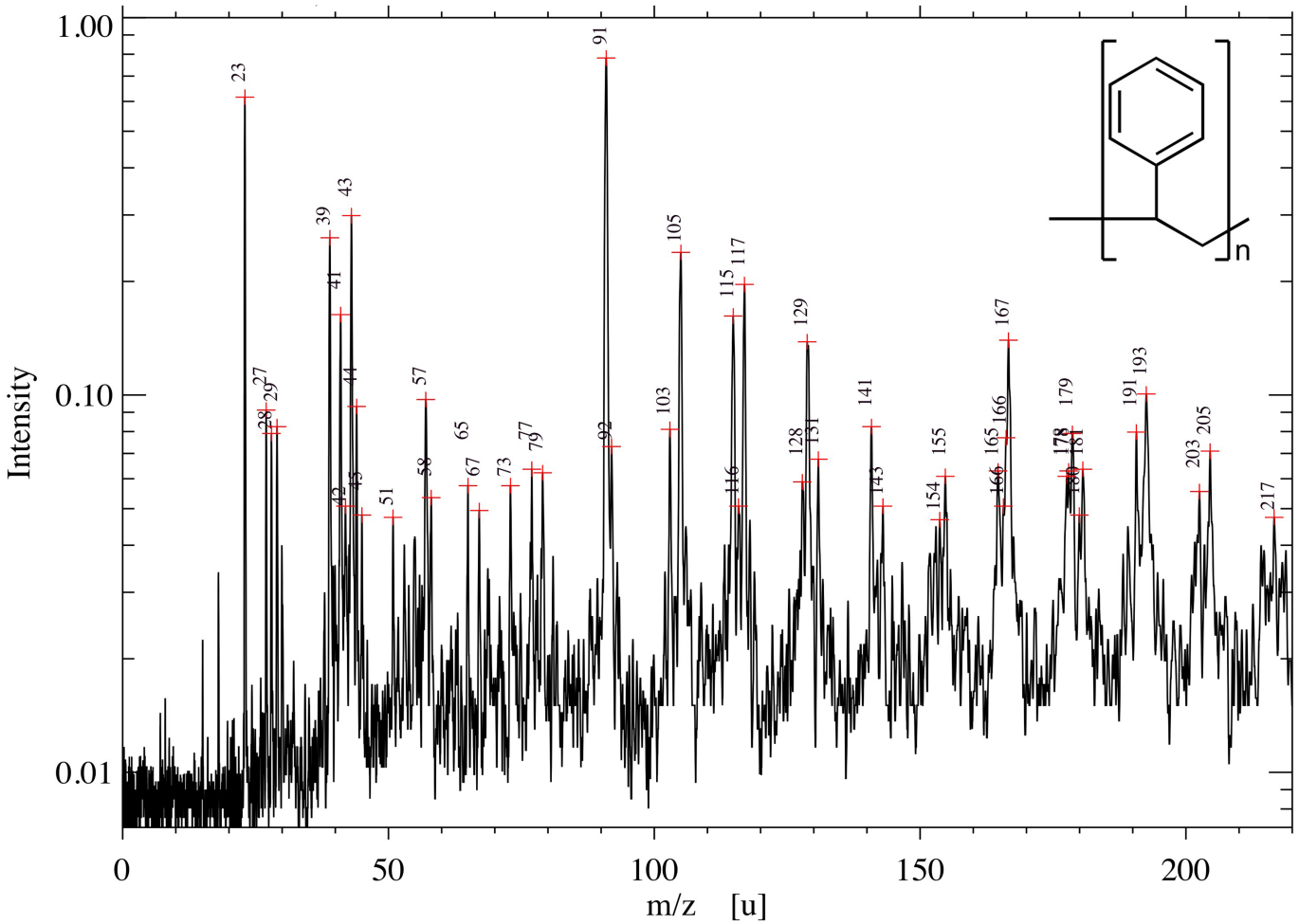
Data availability. All CDA data used for this analysis are archived on PDS-SBN (http://sbn.psi.edu/archive/cocda). The exact time stamps of each data point are listed in Extended Data Table 1.

31. Srama, R. et al. The Cassini cosmic dust analyzer. *Space Sci. Rev.* **114**, 465–518 (2004).
32. Hillier, J. K., McBride, N., Green, S. F., Kempf, S. & Srama, R. Modelling CDA mass spectra. *Planet. Space Sci.* **54**, 1007–1013 (2006).
33. Postberg, F. et al. Composition of jovian dust stream particles. *Icarus* **183**, 122–134 (2006).
34. Srama, R. et al. The cosmic dust analyzer onboard Cassini: ten years of discoveries. *CEAS Space Jour.* **2**, 3–16 (2011).
35. Fiege, K. et al. Calibration of relative sensitivity factors for impact ionization detectors with high-velocity silicate microparticles. *Icarus* **241**, 336–345 (2014).
36. Mocker, A. et al. A 2 MV Van de Graaff accelerator as a tool for planetary and impact physics research. *Rev. Sci. Instrum.* **82**, 095111 (2011).
37. Delcorte, A., Segda, B. G. & Bertrand, P. ToF-SIMS analyses of polystyrene and dibenzanthracene: evidence for fragmentation and metastable decay processes in molecular secondary ion emission. *Surf. Sci.* **381**, 18–32 (1997).
38. Stephan, T., Jessberger, E. K., Heidd, C. H. & Rost, D. TOF-SIMS analysis of polycyclic aromatic hydrocarbons in Allan Hills 84001. *Meteorit. Planet. Sci.* **38**, 109–116 (2003).
39. Wiederschein, F., Vöhringer-Martinez, E. & Postberg, F. Charge separation and isolation in strong water droplet impacts. *Phys. Chem. Chem. Phys.* **17**, 6858–6864 (2015).
40. Le Roy, L. et al. COSIMA calibration for the detection and characterization of the cometary solid organic matter. *Planet. Space Sci.* **105**, 1–25 (2015).
41. Waite, J. H. Jr et al. The Cassini ion and neutral mass spectrometer (INMS) investigation. *Space Sci. Rev.* **114**, 113–231 (2004).
42. Perry, M. E. et al. Cassini INMS measurements of Enceladus plume density. *Icarus* **257**, 139–162 (2015).
43. Kieffer, S. W. et al. A clathrate reservoir hypothesis for Enceladus' south polar plume. *Science* **314**, 1764–1766 (2006).
44. Ingersoll, A. P. & Pankine, A. A. Subsurface heat transfer on Enceladus: conditions under which melting occurs. *Icarus* **206**, 594–607 (2010).
45. Gaston, C. J. et al. Unique ocean-derived particles serve as a proxy for changes in ocean chemistry. *J. Geophys. Res.* **116**, D18310 (2011).
46. Keene, W. C. et al. Chemical and physical characteristics of nascent aerosols produced by bursting bubbles at a model air–sea interface. *J. Geophys. Res.* **112**, D21202 (2007).
47. Facchini, M. C. et al. Primary submicron marine aerosol dominated by insoluble organic colloids and aggregates. *Geophys. Res. Lett.* **35**, L17814 (2008).
48. Russell, L. M. et al. Carbohydrate-like composition of submicron atmospheric particles and their production from ocean bubble bursting. *Proc. Natl Acad. Sci. USA* **107**, 6652–6657 (2010).
49. Burrows, S. M. et al. A physically based framework for modeling the organic fractionation of sea spray aerosol from bubble film Langmuir equilibria. *Atmos. Chem. Phys.* **14**, 13601–13629 (2014).
50. Jayarathne, T. et al. Enrichment of saccharides and divalent cations in sea spray aerosol during two phytoplankton blooms. *Environ. Sci. Technol.* **50**, 11511–11520 (2016).
51. Schmitt-Kopplin, P. et al. Dissolved organic matter in sea spray: a transfer study from marine surface water to aerosols. *Biogeosciences* **9**, 1571–1582 (2012).
52. McCollom, T. M. et al. The influence of carbon source on abiotic organic synthesis and carbon isotope fractionation under hydrothermal conditions. *Geochim. Cosmochim. Acta* **74**, 2717–2740 (2010).
53. Milesi, V. et al. Thermodynamic constraints on the formation of condensed carbon from serpentinization fluids. *Geochim. Cosmochim. Acta* **189**, 391–403 (2016).
54. Williams, L. B. et al. Organic molecules formed in a "primordial womb". *Geology* **33**, 913–916 (2005).
55. Cody, G. D. et al. Establishing a molecular relationship between chondritic and cometary organic solids. *Proc. Natl Acad. Sci. USA* **108**, 19171–19176 (2011).
56. Stevenson, F. J. *Humus Chemistry: Genesis, Composition, Reactions* (Wiley-Interscience, New York, 1982).
57. Hänninen, K. I., Klöcking, R. & Helbig, B. Synthesis and characterization of humic acid-like polymers. *Sci. Total Environ.* **62**, 201–210 (1987).
58. Alexander, C. M. O'D. et al. The nature, origin and modification of insoluble organic matter in chondrites, the major source of Earth's C and N. *Chem. Erde* **77**, 227–256 (2017).
59. Alexander, C. M. O'D., Fogel, M., Yabuta, H. & Cody, G. D. The origin and evolution of chondrites recorded in the elemental and isotopic compositions of their macromolecular organic matter. *Geochim. Cosmochim. Acta* **71**, 4380–4403 (2007).
60. Derenne, S. & Robert, F. Model of molecular structure of the insoluble organic matter isolated from Murchison meteorite. *Meteorit. Planet. Sci.* **45**, 1461–1475 (2010).
61. Remusat, L. Organic material in meteorites and the link to the origin of life. *BIO Web Conf.* **2**, 03001 (2014).
62. Sephton, M. Organic compounds in carbonaceous meteorites. *Nat. Prod. Rep.* **19**, 292–311 (2002).
63. Pizzarello, S., Cooper, G. W. & Flynn, G. J. in *Meteorites and the Early Solar System II* (eds Lauretta D. & McSween H. Y.) 625–651 (Univ. of Arizona Press, Tucson, 2006).
64. Bardyn, A. et al. Carbon-rich dust in comet 67P/Churyumov–Gerasimenko measured by COSIMA/Rosetta. *Mon. Not. R. Astron. Soc.* **469**, S712–S722 (2017).
65. Altwegg, K. et al. Organics in comet 67p – a first comparative analysis of mass spectra from ROSINA–DFMS, COSAC and Ptolemy. *Mon. Not. R. Astron. Soc.* **469**, S130–S141 (2017).
66. Tissot, B. P. & Welte, D. H. *Petroleum Formation and Occurrence* 2nd edn (Springer, Berlin, 1984).
67. Didyk, B. M. & Simoneit, B. R. T. Hydrothermal oil of Guaymas Basin and implications for petroleum formation mechanisms. *Nature* **342**, 65–69 (1989).
68. McKinnon, W. B., Simonelli, D. P. & Schubert, G. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 295–343 (Univ. of Arizona Press, Tucson, 1997).
69. Sephton, M. A. et al. Hydropyrolysis: a new technique for the analysis of macromolecular material in meteorites. *Planet. Space Sci.* **53**, 1280–1286 (2005).
70. Kissel, J. & Krueger, F. R. The organic component in dust from comet Halley as measured by the PUMA mass spectrometer on board Vega 1. *Nature* **326**, 755–760 (1987).
71. Schoell, M. Multiple origins of methane in the Earth. *Chem. Geol.* **71**, 1–10 (1988).
72. Von Damm, K. L. et al. The Escanaba Trough, Gorda Ridge hydrothermal system: temporal stability and subseafloor complexity. *Geochim. Cosmochim. Acta* **69**, 4971–4984 (2005).
73. Horányi, M., Juhász, A. & Morfill, G. E. Large-scale structure of Saturn's E-ring. *Geophys. Res. Lett.* **35**, L04203 (2008).
74. Hsu, H.-W. et al. Understanding the E-ring puzzle. *AGU Fall General Assembly*, abstr. P33E-01 (2016).
75. Waite, J. H. et al. The process of tholin formation in Titan's upper atmosphere. *Science* **316**, 870–875 (2007).
76. Lavvas, P. et al. Aerosol growth in Titan's ionosphere. *Proc. Natl Acad. Sci. USA* **110**, 2729–2734 (2013).
77. Charvat, A. & Abel, B. How to make big molecules fly out of liquid water: applications, features and physics of laser assisted liquid phase dispersion mass spectrometry. *Phys. Chem. Chem. Phys.* **9**, 3335–3360 (2007).
78. Srama, R. *Cassini–Huygens and Beyond—Tools for Dust Astronomy*. Habil. Thesis, Univ. of Stuttgart (2009).

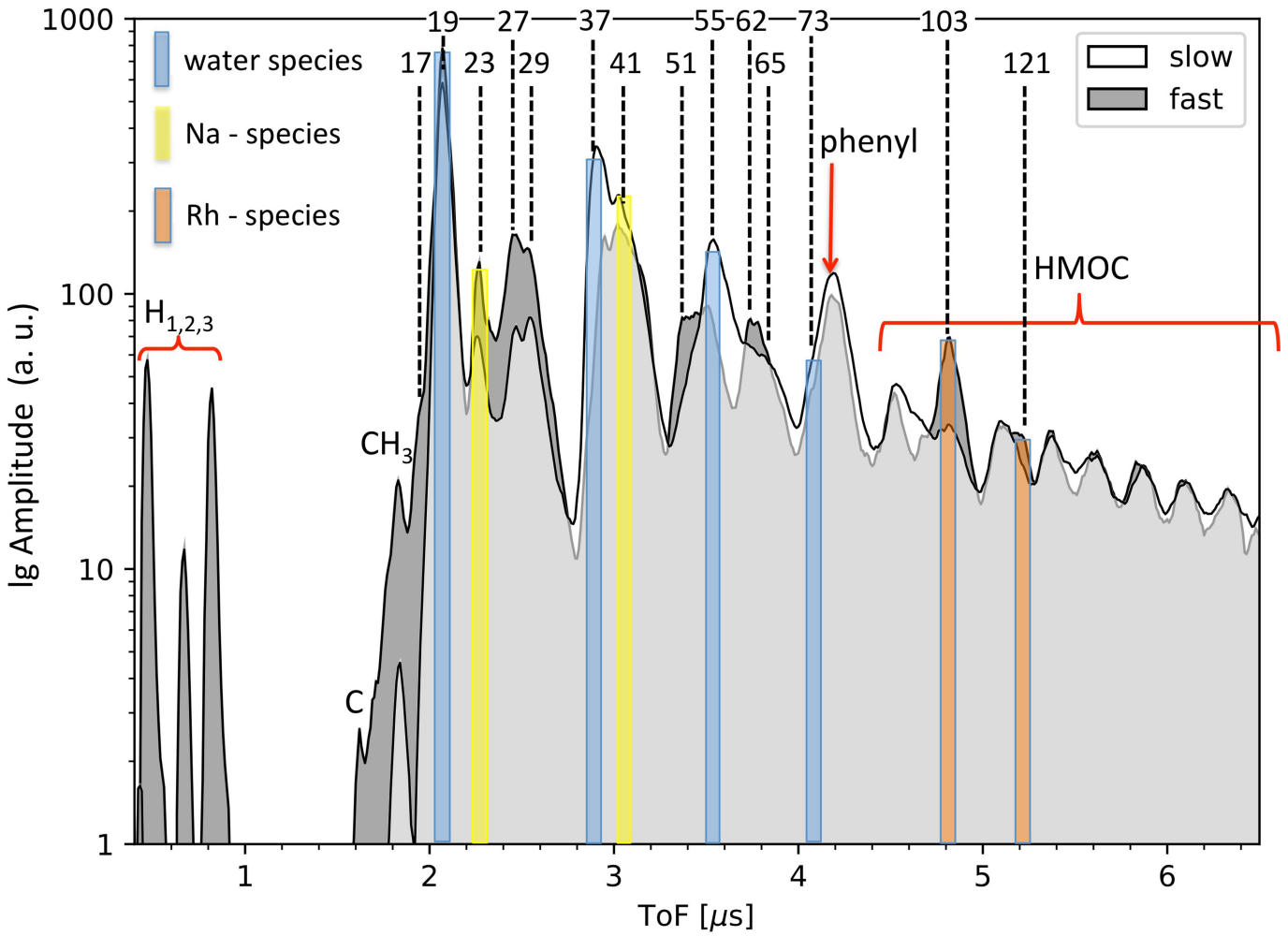


Extended Data Fig. 1 | See next page for caption.

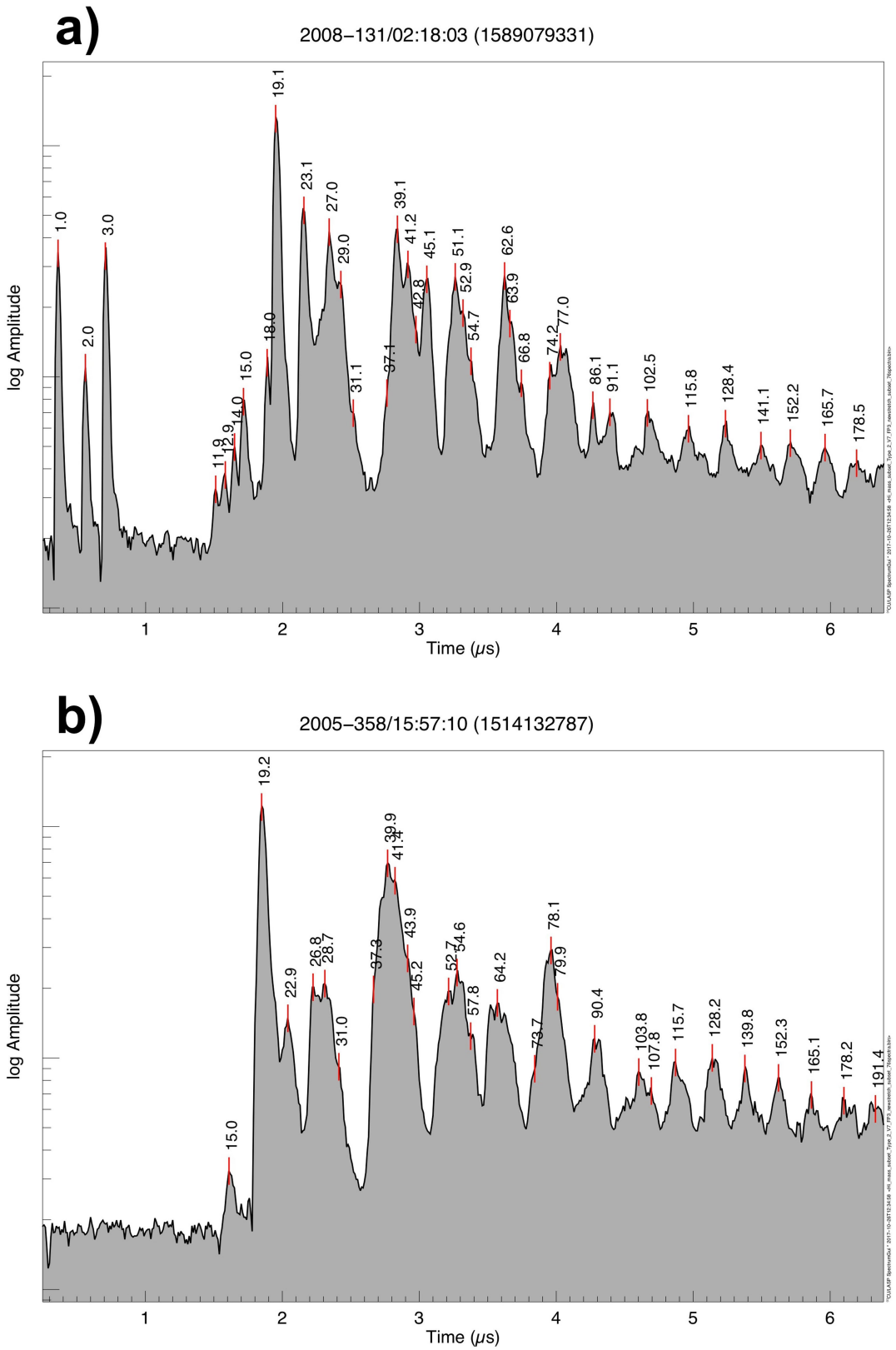
Extended Data Fig. 1 | Semi-quantitative display of CDA HMOC spectra. a, **b**, Identified organic mass lines of individual HMOC spectra. The distribution of resolved mass lines and flank peaks of 64 HMOC spectra with the most distinct HMOCs are shown (see Methods, 'Selection of 64 high-quality spectra for Fig. 1 and Extended Data Figs. 1 and 3'). 19 more spectra with a high level of interference with water-cluster ions or low signal-to-noise ratio are not included here (see Extended Data Table 1). All peaks depicted here are also part of the data shown in Fig. 1b. The spectrum number (as defined in Extended Data Table 1) is indicated on the left as an identifier of the event. The extent is indicated by the horizontal length and the relative normalized amplitude of each spectral feature is given by the length in the vertical direction and the colour code (red being the highest and blue the lowest amplitude). The largest horizontal span of the symbol marks the peak maximum. In **b**, the amplitudes between 70 u and 85 u shown in grey indicate that they are not to scale with the symbols shown at higher masses (they would be much larger; see Fig. 1b for comparison). Spectra are sorted by their impact speed, as estimated from the orbital elements of the impacting grain, with the highest speed (~15 km s^{-1}) at the top of the graphs and the lowest (~5 km s^{-1}) at the bottom. Because the exact orbital elements are unknown, each impact speed has substantial intrinsic uncertainties, given in Extended Data Table 1. The 12 spectra for which a minimum impact speed could be derived from the presence of hydrogen mass lines (Extended Data Table 1; see Methods, 'Selection of 64 high-quality spectra for Fig. 1 and Extended Data Figs. 1 and 3') are placed at the top. The highest mass at which the recording of the CDA TOF spectrum ends varies between 174 u and 226 u (Methods, 'Short description of CDA's chemical analyser subsystem'), as indicated by the grey horizontal bars. As a consequence, the frequency of the HMOC peaks around 178 u and 191 u in Fig. 1b is reduced because not all individual spectra cover this mass range. The absolute masses in each individual spectrum have an intrinsic uncertainty (absolute value) of ±1 u at 80 u and ±2 u at 180 u owing to the limited calibration accuracy of the CDA in this high-mass regime. The mass intervals between peaks, however, are accurate to the integer level.



Extended Data Fig. 2 | Impact ionization laboratory spectrum of a polystyrene bead. The figure was modified from figure 5 of ref. [19]. The *x*-axis shows *m* (mass) over *z* (cation charge), with $z = 1$ for all major species. The impact ionization TOF mass spectrum of a polystyrene particle with a radius of ~1 μm was recorded at the Heidelberg dust accelerator facility[36] with an impact speed of 5.2 km s^{-1}. Above 100 u and below 70 u the spectrum shows cationic fragments, in good agreement with the CDA HMOC spectra and their characteristic spacing of 12.5 u. The inset shows the molecular structure of the polymer. See the main text and Methods section 'Inferring the origin of HMOC peaks in CDA spectra' for further discussion.

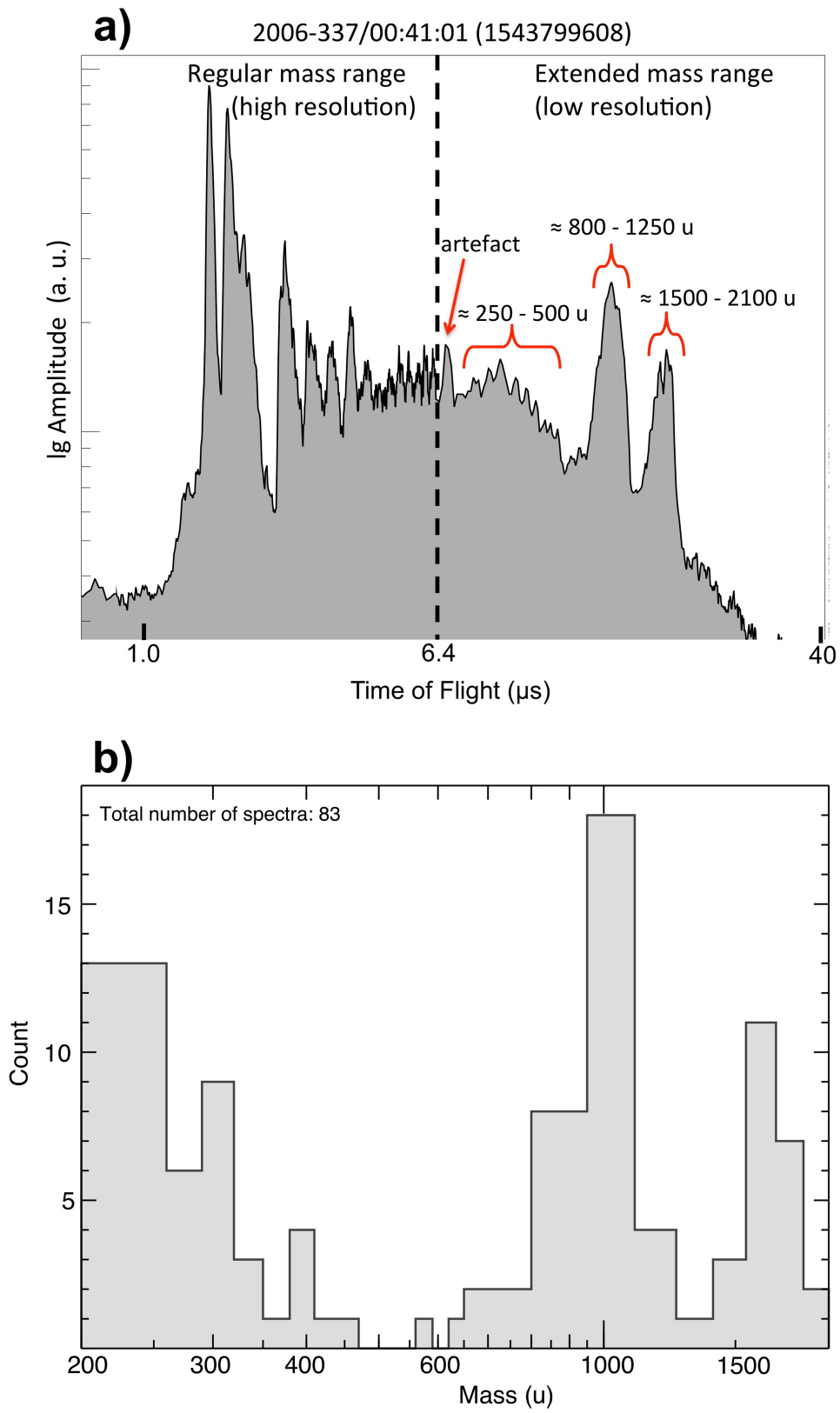


Extended Data Fig. 3 | Comparison of CDA HMOC spectra from fast and slow impacts. White and grey spectra represent the average of all spectra from impacts below and above ~10 km s^{-1}, respectively (see Methods, 'Selection of 64 high-quality spectra for Fig. 1 and Extended Data Figs. 1 and 3'). All signatures with possible major contributions from inorganic species are colour-shaded as in Fig. 1. Signatures not marked are exclusively or mostly due to organic cations. The abundance and position of the HMOC species is relatively independent of the impact speed of the ice grain (see also Extended Data Fig. 1). By contrast, fast impacts induce stronger organic fragmentation signatures at masses below 70 u and HMOCs form more distinct, evenly spaced groups, characteristic of impact-induced dissociation processes. In turn, slow impacts show more abundant intact benzene-like cations. There seems to be a tendency of some organic cations to carry fewer H atoms at fast impacts (27 u, 51 u and 62 u), which is indicative of 'softer' ionization from the slower impact. In fast spectra, interference with water-cluster ions is less frequent than at lower speeds. In contrast to fast spectra, fragmentation below CH_3^+ (15 u) is usually not observed in slow spectra. Spectra from slow impacts are prone to abundant water clustering, creating mass lines of the form $H^+(H_20)_n$, with $n = 1$–4, at 19 u, 37 u, 55 u and 73 u (blue). In fast spectra, clustering is limited and only the mass lines at 19 u and 37 u are generally present; occasionally, formation of the smaller water ions OH^+ (17 u) and H_2O^+ (18 u) is observed. Similarly, Rh^+ (103 u) forms from excavation of the impact target only at fast impacts and interferes with HMOC species there. To a lesser extent, this is also true for the rhodium–water cluster $Rh^+(H_2O)$ at 121 u. See the individual CDA spectra in Extended Data Fig. 4 for comparison. a.u., arbitrary units; ToF, time of flight; lg, log.



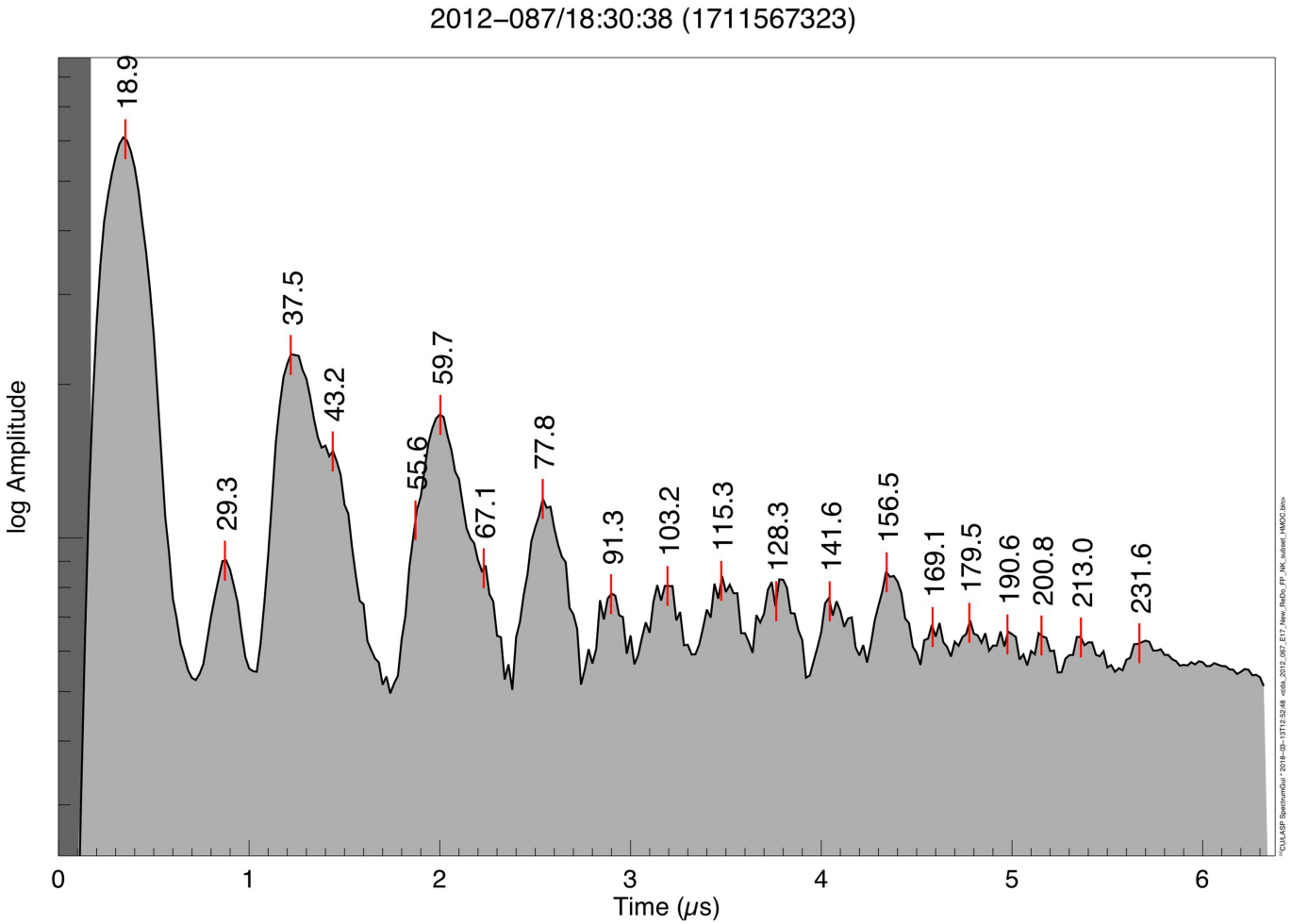
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Example CDA spectra from individual HMOC-type ice grains. In these individual spectra, the peak definition is naturally higher than in the co-added spectra shown in Fig. 1 and Extended Data Fig. 3, and therefore some of the spectral features collected in Fig. 1b become more apparent. **a**, HMOC spectrum from one of the fastest recorded impacts (12–18 km s^{-1}). The appearance of hydrogen cations (H^+, H_2^+ and H_3^+) at 1 u, 2 u and 3 u, as well as the disintegration of the CH_3^+ ion into CH_2^+, CH^+ and C^+ (12 u–15 u) and the formation of H_2O^+ (18 u), are evidence of the high-speed impact. The abundance of unsaturated small cations below 70 u, probably fragments from aromatic structures, is increased compared to slower spectra. The frequently occurring mass line at 45 u (Fig. 1b and Extended Data Fig. 1) is noticeable; it cannot originate from pure hydrocarbons and requires heteroatoms, probably oxygen in this case. While a 45 u feature is quite common in our HMOC dataset, the peak at 86 u is only apparent in this spectrum. **b**, HMOC spectrum from a grain detected at intermediate speed (5–8 km s^{-1}). High-mass fragments and benzene species are abundant whereas further fragmentation of the benzene ring into C_5 and C_4 species is less apparent compared to high-velocity impacts (**a**). We note that organic cations with 2, 3, 4 and 5 C atoms show a tendency to carry more H atoms compared with the high-speed impact, which is indicative of 'softer' ionization from the slower impact. Organic fragmentation below CH_3^+ is usually not observed in this speed regime.

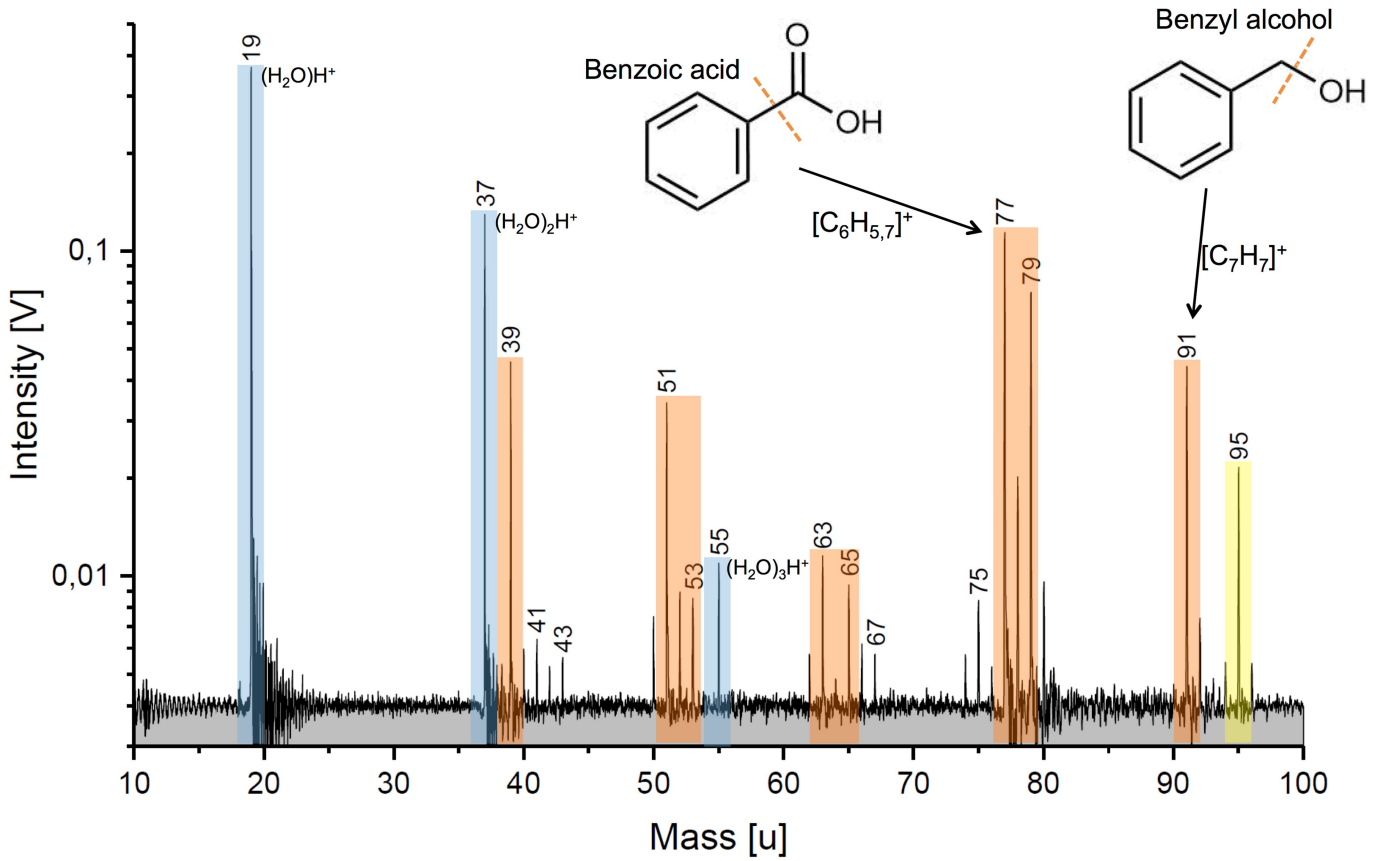


Extended Data Fig. 5 | See next page for caption.

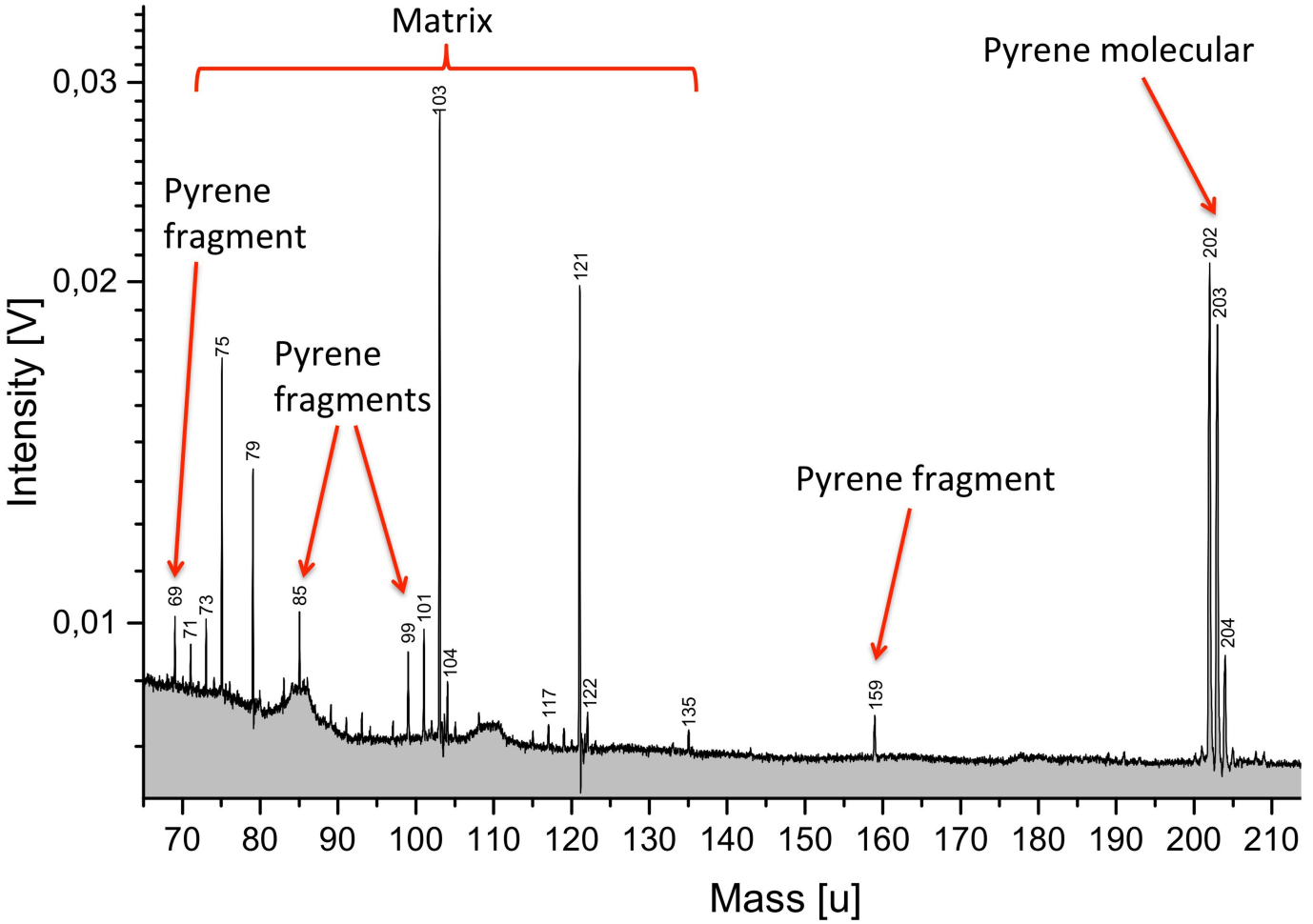
Extended Data Fig. 5 | Example HMOC CDA mass spectrum with extended mass range and statistics for all features. a, Ice grain spectrum showing the HMOC event with the strongest extended mass range signal of the dataset. The dashed line at 6.4 μs divides the spectrum into the high-resolution part (10 ns sampling) and the low-resolution part (100 ns sampling) (see Methods, 'Short description of CDA's chemical analyser subsystem'). There are several relatively narrow peaks between 250 u and 500 u and two much more extended features peaking at about 1,000 u and 1,800 u. In this case, the cations with mass in excess of 200 u are more than twice as abundant (defined by the area under the curve) as those below 200 u. We note the logarithmically scaled TOF axis in this case. These features are usually less frequent and less pronounced than in the extreme case shown here. The extended spectrum frequently shows an instrument-artefact peak at 6.8 μs, which was not considered in our analysis. **b**, Histogram showing the frequency of occurrences of the features observed in the extended mass range. The definition, and thus significance, of peaks in the extended spectrum is generally lower than in the nominal spectrum. In particular, features above 500 u are sometimes ambiguous and their interpretation should be taken with great caution. However, the statistics shows three preferred mass regions: 200 u–500 u with decreasing frequency, around 1,000 u and around 1,700 u. Even if no sizeable peaks are present, the cation signal in the HMOC spectra is generally higher than the noise level when the low-resolution recording starts and typically only decays to noise level at around 500 u or later.



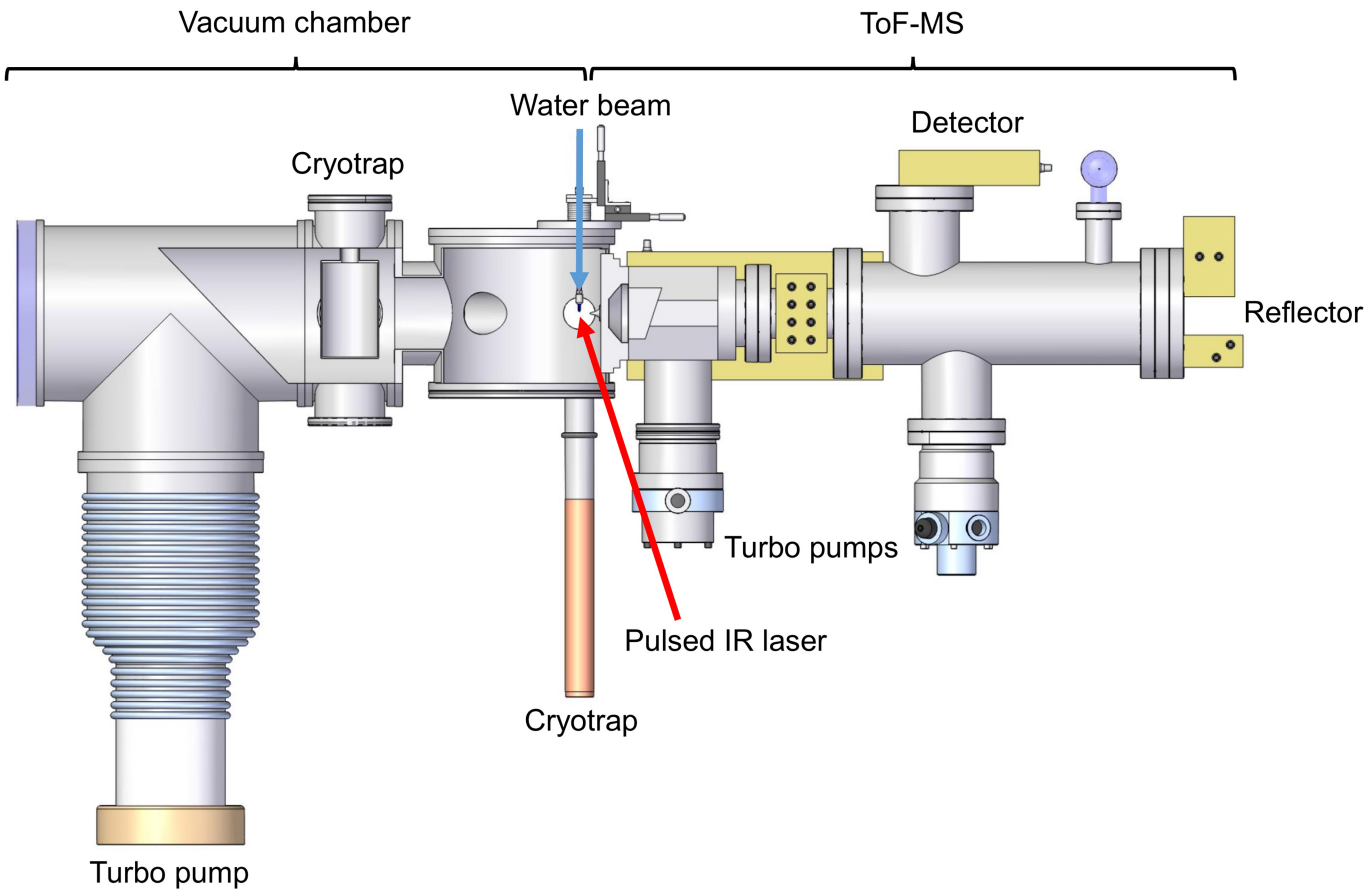
Extended Data Fig. 6 | CDA HMOC spectrum recorded in the Enceladean plume. During Cassini's E17 flyby of Enceladus' south pole at 75 km altitude, where the CDA recorded about 40 plume spectra with its full mass range, one spectrum was of the HMOC type. This is in agreement with the proportion of this particle type being a few per cent in the plume and in the E ring close to Enceladus (see Methods, 'Relative frequency of HMOC-type grains depends on impact speed and distance to Enceladus orbit.'). The flyby speed determined the impact speed of 8.6 km s^{-1} and the particle had a radius of about 2 μm. To operate the CDA in the dense dust environment of the plume, the instrument settings had to be modified in a way that compromised the spectrum quality (lower sensitivity and lower mass resolution; see Extended Data Fig. 4 for comparison). The spectrum is baseline-corrected. On the only occasion when the CDA recorded a large number of spectra with high cadence directly in the plume during Cassini's E5 flyby[9], the spectral range was truncated below about 100 u to allow for a higher data rate. This unfortunately did not allow the identification of the defining HMOC signatures.



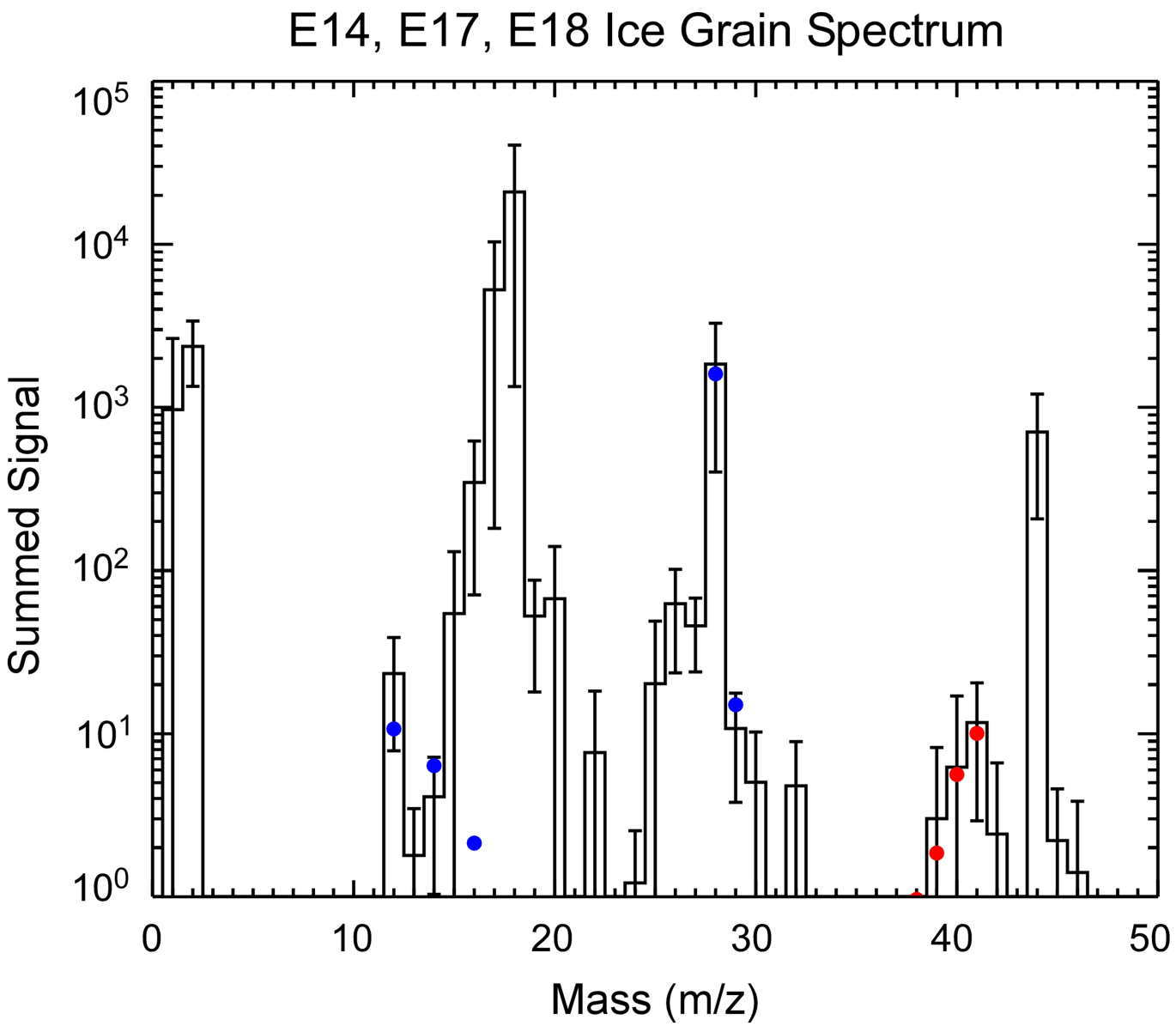
Extended Data Fig. 7 | Laser ionization mass spectrum of benzoic acid and benzyl alcohol dissolved in water. Analogue TOF mass spectrum recorded with the liquid microbeam ionization setup (see Methods, 'Laser dispersion analogue experiments for icy dust impacts' and Extended Data Fig. 9) to simulate the formation of tropylium and benzene cations and their fragmentation ions at impact speeds[78] of the order of 10 km s^{-1}. The concentrations of benzoic acid and benzyl alcohol are 3 g l^{-1} and 0.2 g l^{-1}, respectively. Water ions are marked in blue, aromatic ions and ions from aromatic fragmentation are marked in orange and mixed organic–water species are yellow. To yield both benzene cations (77 u–79 u) and tropylium ions (91 u), two different aromatic structures are required (Fig. 2). The predominant aromatic fragments of benzoic acid are at 77 u and 79 u, whereas benzyl alcohol almost exclusively forms tropylium ions at 91 u. The peak at 95 u is a water cluster of the phenyl cation, which is much more pronounced than in the HMOC spectra. Although the strong phenyl–water cluster signature here illustrates the intimate mixing of organics with water, the much lower 95 u signature in HMOC spectra argues for less efficient mixing of organics with water there, probably due to a core–shell structure that physically separates organics from ice in the grain. Cations from the fragmented ring can be seen at 39 u, 51 u–53 u and 63 u–65 u and agree with the CDA observations (Fig. 1b). In contrast to the CDA spectra, however, saturated C_3 fragments (41 u–43 u) are depleted, and C_2 (27 u–29 u) and C_1 (15 u) fragment cations are entirely absent, confirming the presence of an abundance of aliphatic cations in HMOC grains. The ratios of benzene and tropylium ions and the water ions match the HMOC spectra well. The total concentration of organic species used here (~0.32% by weight) can be used to estimate a lower limit for the concentration of organics in CDA HMOC grains for two reasons. First, in the analogue experiment we selected substances that most efficiently yield the desired aromatic species and other, less efficient precursors would yield even lower signals at 77 u and 91 u. Second, to account for both the low- and high-mass fragments between 100 u and 2,000 u, which are absent in the laboratory spectrum, additional organic substances or larger molecules would be needed to further increase the organic concentration. Therefore, the concentration in Enceladean HMOC ice grains in many cases can be estimated to be near or even above the per cent level.



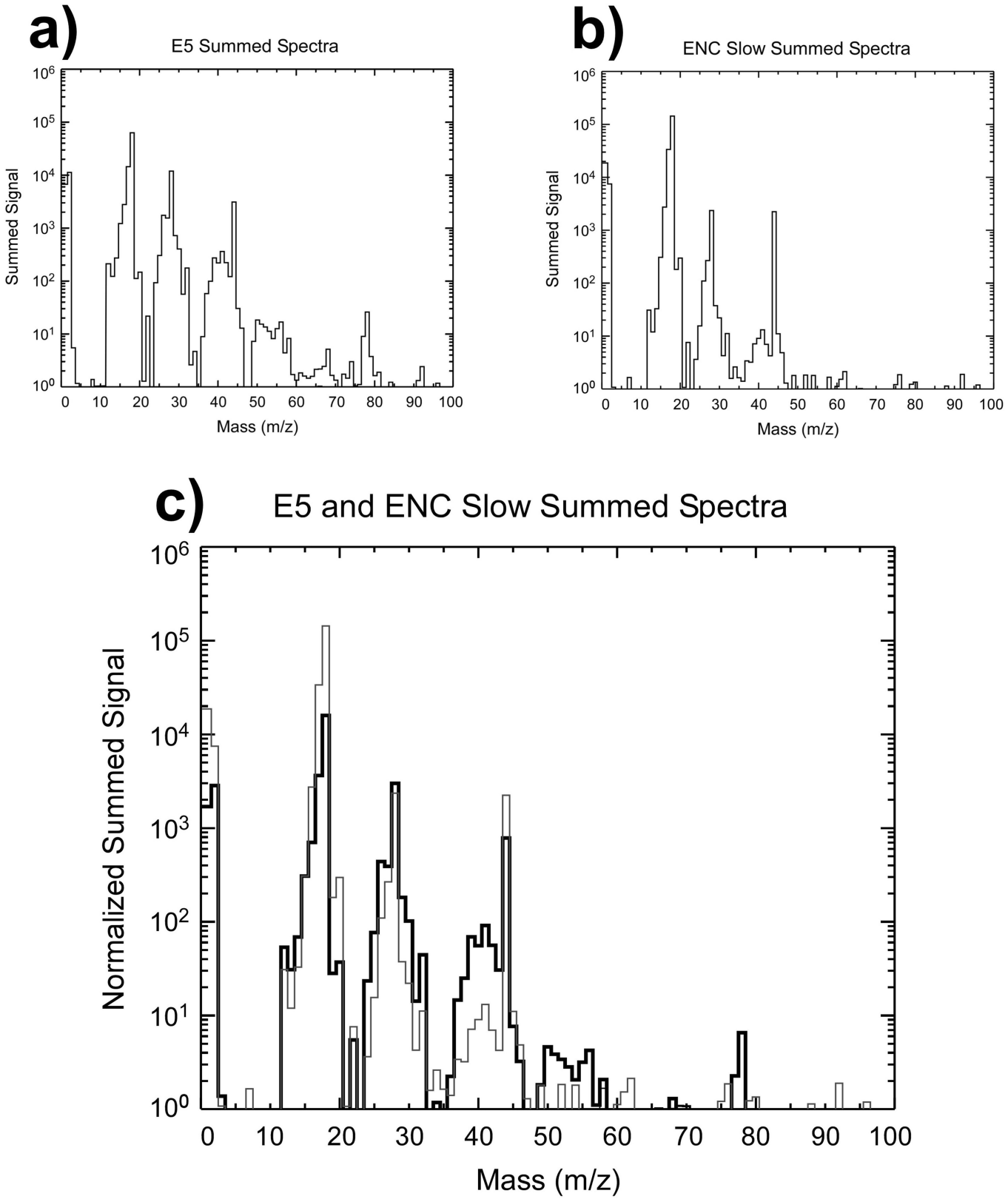
Extended Data Fig. 8 | Laser ionization analogue spectrum of pyrene in water–acetic acid mixture. Cationic TOF mass spectrum recorded with the liquid microbeam ionization setup (see Methods, 'Laser dispersion analogue experiments for icy dust impacts', and Extended Data Fig. 9) containing 0.1% pyrene dissolved in a mixture of water and acetic acid. The laser pulse simulates an ice grain impact with a speed[78] of about 10 km s^{-1}. Features marked as 'pyrene fragment' do not appear in the blank experiment with just the solvent mixture and are either direct pyrene fragments or cations formed from pyrene fragments clustering with the solvent (for example, at 159 u). The molecular mass lines of pyrene are about 10 times more abundant than those of any fragments, indicating the stability of the PAH molecule. In contrast to CDA HMOC spectra, no isolated benzene ring fragment (77 u or 91 u) forms.



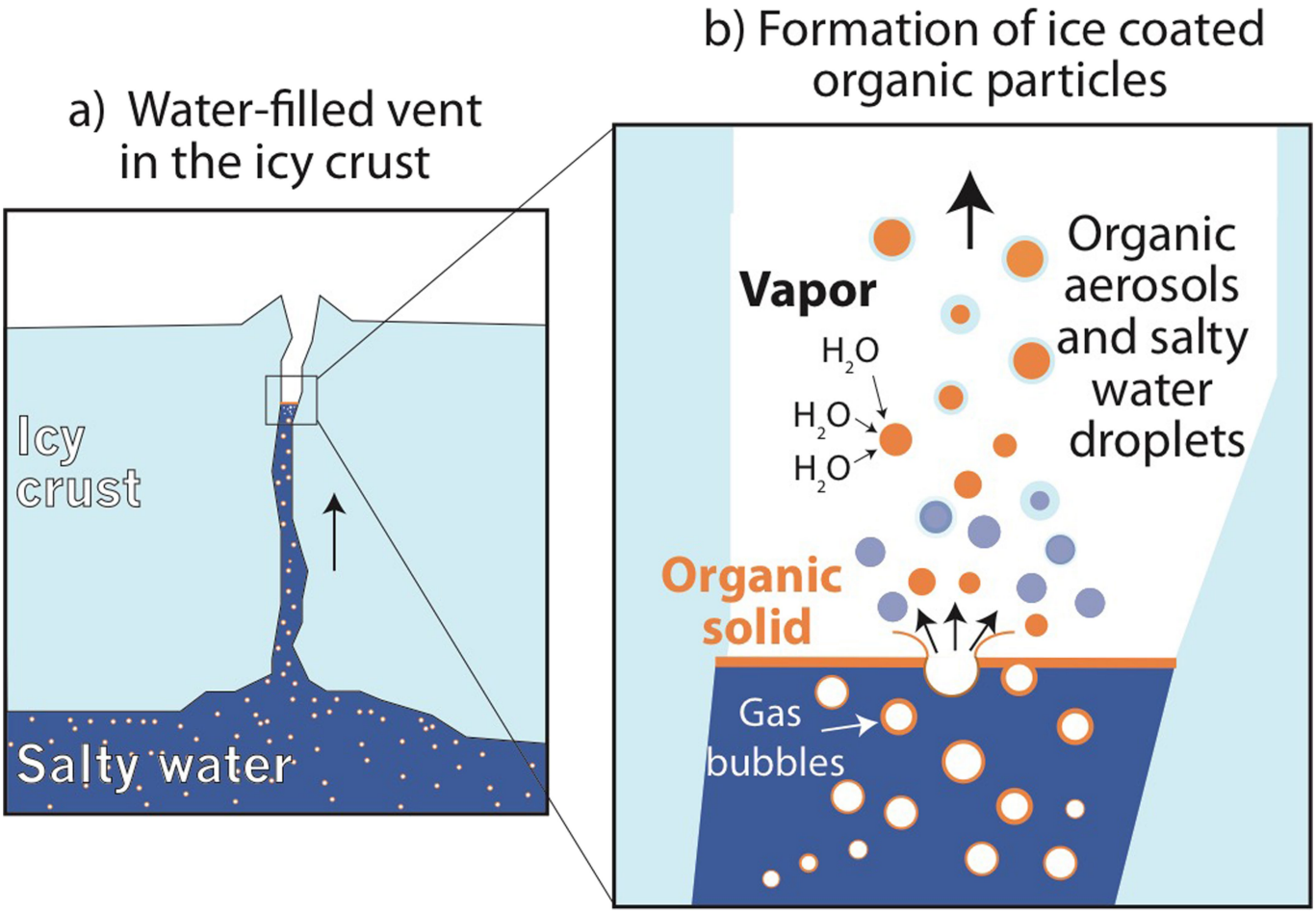
Extended Data Fig. 9 | Laboratory setup used to simulate ice grain impacts onto space-borne impact ionization detectors. See Methods, 'Laser dispersion analogue experiments for icy dust impacts', for a detailed description of the setup.



Extended Data Fig. 10 | Co-added INMS ice grain spike spectrum from three plume encounters. See Methods section 'INMS ice grain spectrum' for details on how the spectrum was composed and analysed. Error bars (1 s.d.) are derived from the dispersion of the count rate from the three separate measurements of the individual encounters. The spectrum suggests the presence of CO fragments (blue circles) as an oxygen-bearing species. N_2 has very low abundance and contributes less than ~10% of the 28 u signal. CO_2 and C_2H_4 collectively contribute less than ~10% of the 28 u signal. CO (blue circles) is required to fit the rest of the 28 u signal and the entire 29 u signal and matches its other dissociative peaks (C at 12 u and CO^{++} at 14 u) well. The spectrum also indicates the presence of nitrogen-bearing species: the 'stair step' pattern around 41 u matches best to the C_2H_3N spectrum (red circles).



Extended Data Fig. 11 | INMS spectra used to produce the differenced spectrum in Fig. 3. a, **b**, The individual spectra for fast (E5; **a**) and slow (**b**) flybys are shown. **c**, The spectra of **a** and **b** are plotted with the E5 (black) spectrum, which is normalized to match the 15 u signal of the slow spectrum (grey). The residual (difference) between the two spectra is plotted in Fig. 3.



Extended Data Fig. 12 | Schematic on the formation of organic condensation cores from a refractory organic film. a, Ascending gas bubbles in the ocean[25] efficiently transport organic material[30] into water-filled cracks in the south polar ice crust. **b**, Organics ultimately concentrate in a thin organic layer (orange) on top of the water table, located inside the icy vents. When gas bubbles burst, they form aerosols made of insoluble organic material that later serve as efficient condensation cores for the production of an icy crust from water vapour, thereby forming HMOC-type particles. In parallel, larger, pure salt-water droplets form (blue), which freeze and are later detected by the CDA as salt-rich type-3 ice particles in the plume[8,9].

Extended Data Table 1 | List of 83 CDA mass spectra identified as of HMOC type

#	UTC	SCLK	QI (C)	R (R_S)	Z (R_S)	V_{imp}	$V_{-\sigma}$	$V_{+\sigma}$	Radius (μm)		
1	**2004-302/21:31:38**	1477691821	1.34E-14	8.6	-0.22	6.4	4.7	8.0	0.8	0.3	2
2	**2005-068/12:11:47**	1489063103	4.04E-14	3.53	0.01	5.0	3.7	6.3	1.2	0.5	3.2
3	**2005-068/19:12:50**	1489088367	2.30E-14	6.24	0.02	8.4	6.2	10.6	0.6	0.2	1.5
4	2005-068/19:57:18	1489091035	5.29E-14	6.61	0.02	8.4	6.2	10.6	0.7	0.3	1.7
5	**2005-177/09:15:19**	1498470176	1.02E-13	5.68	-1.44	8.6*	6.4	10.9	0.7	0.3	1.9
6	**2005-177/09:16:46**	1498470263	4.43E-14	5.67	-1.44	8.6	6.4	10.9	0.6	0.3	1.6
7	**2005-177/10:14:42**	1498473739	1.68E-13	5.21	-1.13	8.7	6.4	11.0	0.7	0.3	2
8	**2005-177/10:15:39**	1498473796	2.55E-14	5.2	-1.13	8.7	6.4	11.0	0.6	0.2	1.5
9	2005-177/10:19:23	1498474020	1.32E-14	5.18	-1.11	8.7	6.4	11.0	0.5	0.2	1.4
10	**2005-177/10:23:56**	1498474293	1.62E-14	5.14	-1.08	8.7	6.4	11.0	0.5	0.2	1.4
11	**2005-177/11:24:38**	1498477935	2.12E-14	4.69	-0.74	8.6	6.4	10.9	0.6	0.2	1.5
12	**2005-177/11:36:22**	1498478639	2.88E-14	4.61	-0.67	8.6	6.4	10.9	0.6	0.2	1.5
13	2005-177/11:46:21	1498479238	1.40E-14	4.54	-0.62	8.6	6.4	10.9	0.5	0.2	1.4
14	**2005-177/12:00:40**	1498480097	5.02E-14	4.45	-0.53	8.6*	6.3	10.8	0.6	0.3	1.7
15	2005-177/12:22:55	1498481432	5.43E-14	4.3	-0.4	8.5*	6.3	10.7	0.6	0.3	1.7
16	**2005-177/12:24:48**	1498481545	2.54E-14	4.29	-0.39	8.5*	6.3	10.7	0.6	0.2	1.5
17	**2005-177/12:31:51**	1498481968	3.26E-14	4.25	-0.35	8.5	6.3	10.7	0.6	0.3	1.6
18	**2005-177/12:48:40**	1498482977	1.10E-14	4.15	-0.25	8.4	6.2	10.6	0.5	0.2	1.4
19	**2005-177/12:49:08**	1498483005	4.12E-14	4.15	-0.24	8.4	6.2	10.6	0.6	0.3	1.7
20	**2005-177/12:50:33**	1498483090	1.82E-14	4.14	-0.24	8.4	6.2	10.6	0.6	0.2	1.5
21	**2005-177/12:56:32**	1498483449	5.27E-14	4.11	-0.2	8.3	6.2	10.5	0.7	0.3	1.8
22	**2005-177/15:48:06**	1498493743	2.20E-14	3.61	0.81	6.9	5.1	8.7	0.7	0.3	1.9
23	**2005-177/16:47:37**	1498497314	3.55E-13	3.69	1.11	6.6	4.9	8.3	1.2	0.5	3.1
24	2005-267/03:21:33	1506224999	3.40E-14	5.24	0.01	9.1	6.8	11.5	0.6	0.2	1.5
25	**2005-267/03:22:22**	1506225048	9.24E-14	5.25	0.01	9.2	6.8	11.5	0.6	0.3	1.7
26	**2005-267/03:35:30**	1506225836	2.75E-14	5.37	0.01	9.2	6.8	11.6	0.5	0.2	1.4
27	**2005-267/07:34:31**	1506240177	2.94E-14	7.52	0	8.8	6.5	11.2	0.6	0.2	1.5
28	**2005-303/04:28:34**	1509339440	3.74E-14	5.76	0.04	6.4	4.8	8.1	0.9	0.4	2.3
29	2005-303/04:43:14	1509340320	1.75E-14	5.85	0.04	6.5	4.8	8.2	0.8	0.3	2
30	**2005-303/04:48:04**	1509340610	4.42E-14	5.88	0.04	6.6	4.8	8.3	0.9	0.4	2.3
31	**2005-303/05:03:27**	1509341533	4.64E-14	5.98	0.04	6.6	4.9	8.4	0.9	0.4	2.3
32	2005-303/05:43:23	1509343929	4.32E-14	6.24	0.04	6.8	5.0	8.6	0.8	0.3	2.2
33	2005-303/05:53:09	1509344515	5.82E-13	6.3	0.04	6.8	5.1	8.6	1.2	0.5	3.2
34	**2005-303/05:59:07**	1509344873	7.83E-14	6.34	0.04	6.9	5.1	8.7	0.9	0.4	2.4
35	**2005-303/06:36:17**	1509347103	3.31E-13	6.59	0.04	7.0	5.2	8.8	1.1	0.4	2.9
36	**2005-303/06:41:04**	1509347390	8.31E-14	6.62	0.04	7.0	5.2	8.8	0.9	0.4	2.3
37	2005-330/22:05:55	1511735696	2.39E-14	8.97	-0.06	7.3	5.4	9.2	0.7	0.3	1.8
38	**2005-358/14:37:41**	1514128018	6.68E-14	6.2	-0.03	6.8	5.0	8.6	0.9	0.4	2.3
39	**2005-358/14:49:12**	1514128709	4.32E-13	6.13	-0.03	6.8	5.0	8.5	1.2	0.5	3.1
40	**2005-358/15:57:10**	1514132787	2.07E-14	5.71	-0.02	6.4	4.7	8.1	0.8	0.3	2.1
41	**2005-358/15:57:39**	1514132816	4.30E-14	5.7	-0.02	6.4	4.7	8.1	0.9	0.4	2.4
42	2005-358/16:09:40	1514133537	7.77E-15	5.63	-0.02	6.3	4.7	8.0	0.7	0.3	1.8
43	2005-359/03:01:39	1514172656	6.11E-13	5.78	0.04	6.5	4.8	8.2	1.3	0.5	3.5
44	2005-359/04:06:40	1514176557	2.87E-14	6.19	0.04	6.8	5.0	8.6	0.8	0.3	2.1
45	**2005-359/04:06:52**	1514176569	9.60E-14	6.19	0.04	6.8	5.0	8.6	0.9	0.4	2.5
46	**2005-359/05:10:40**	1514180397	5.38E-14	6.62	0.04	7.0	5.2	8.8	0.8	0.3	2.2
47	2005-359/05:20:58	1514181015	4.33E-14	6.69	0.05	7.0	5.2	8.9	0.8	0.3	2.1
48	**2005-359/06:01:55**	1514183472	2.19E-14	6.98	0.05	7.1	5.3	9.0	0.7	0.3	1.9
49	**2005-359/06:26:12**	1514184929	7.14E-14	7.15	0.05	7.2	5.3	9.0	0.8	0.3	2.2
50	**2005-359/06:32:15**	1514185292	1.35E-13	7.19	0.05	7.2	5.3	9.0	0.9	0.4	2.4
51	**2006-016/08:08:18**	1516091868	5.75E-14	12.91	-0.08	6.5	4.8	8.2	0.9	0.4	2.4
52	**2006-056/05:15:38**	1519537530	1.26E-14	6.46	-0.03	5.4	4.0	6.8	0.9	0.4	2.4
53	**2006-056/17:46:16**	1519582569	1.65E-14	6.82	0.04	5.7	4.2	7.2	0.9	0.4	2.3
54	**2006-056/18:48:22**	1519586295	9.86E-15	7.15	0.04	6.0	4.4	7.5	0.8	0.3	2
55	2006-056/22:14:58	1519598691	1.32E-14	8.4	0.05	6.5	4.8	8.1	0.7	0.3	1.9
56	**2006-057/02:02:07**	1519612320	6.24E-14	9.88	0.07	6.6	4.9	8.4	0.9	0.4	2.4
57	**2006-057/02:05:38**	1519612531	3.55E-14	9.9	0.07	6.6	4.9	8.4	0.8	0.3	2.2
58	**2006-080/15:29:58**	1521648004	1.64E-14	11.6	0.08	6.7	4.9	8.4	0.7	0.3	1.9
59	2006-080/19:20:46	1521661852	1.79E-14	13.1	0.09	6.5	4.8	8.2	0.8	0.3	2
60	**2006-337/00:41:01**	1543799608	8.17E-14	4.99	0.04	12.6	9.3	15.8	0.4	0.2	1.1
61	**2006-337/00:46:46**	1543799953	4.60E-13	5.01	0.11	12.5*	9.3	15.8	0.5	0.2	1.4
62	**2006-337/00:55:19**	1543800466	4.77E-14	5.04	0.21	1.7	1.3	2.1	0.4	0.2	1.1
63	**2006-337/00:59:05**	1543800692	1.55E-14	5.05	0.25	12.5	9.2	15.8	0.3	0.1	0.9
64	**2006-337/01:18:39**	1543801866	1.86E-14	5.13	0.48	12.4	9.2	15.6	0.4	0.1	0.9
65	**2006-337/01:26:38**	1543802345	3.85E-14	5.16	0.57	12.4*	9.1	15.6	0.4	0.2	1
66	**2006-337/04:59:37**	1543815124	1.57E-13	6.24	2.93	10.6*	7.8	13.3	0.6	0.2	1.5
67	2006-337/05:49:38	1543818125	1.75E-14	6.54	3.43	10.1	7.5	12.8	0.4	0.2	1.2
68	**2006-349/04:52:09**	1544851483	4.86E-14	7.89	-2.66	8.9	6.5	11.2	0.6	0.2	1.6
69	**2006-349/16:06:48**	1544891962	3.95E-14	10.08	2.6	8.3	6.1	10.5	0.6	0.3	1.7
70	**2007-130/19:21:23**	1557518117	8.31E-14	4.8	0.08	10.3*	7.6	13.0	0.5	0.2	1.5
71	**2007-130/19:32:19**	1557518773	5.98E-13	4.75	-0.02	10.3*	7.6	13.0	0.7	0.3	1.9
72	**2007-130/19:37:15**	1557519069	4.10E-14	4.73	-0.06	10.3	7.6	13.0	0.5	0.2	1.3
73	**2007-130/19:44:45**	1557519519	6.15E-14	4.7	-0.13	10.3	7.6	13.0	0.5	0.2	1.4
74	**2007-130/19:50:29**	1557519863	3.34E-14	4.68	-0.18	10.3	7.6	13.0	0.5	0.2	1.3
75	2007-130/19:51:50	1557519944	1.49E-13	4.67	-0.2	10.3	7.6	13.0	0.6	0.2	1.6
76	**2007-130/19:53:50**	1557520064	3.66E-14	4.66	-0.21	10.3*	7.6	12.9	0.5	0.2	1.3
77	**2008-130/23:08:03**	1589067930	1.16E-13	4.67	0.76	14.3*	10.6	18.0	0.4	0.2	1
78	2008-130/23:57:13	1589070880	2.02E-14	4.5	0.1	14.6*	10.8	18.5	0.3	0.1	0.8
79	**2008-131/00:20:35**	1589072282	4.29E-14	4.44	-0.21	14.7*	10.9	18.6	0.3	0.1	0.9
80	**2008-131/02:18:04**	1589079331	5.10E-14	4.3	-1.75	14.1*	10.4	17.8	0.3	0.1	0.9
81	**2008-131/03:00:00**	1589081847	8.66E-14	4.32	-2.26	13.5	10.0	17.0	0.4	0.2	1
82	**2008-131/03:16:31**	1589082838	1.34E-13	4.35	-2.46	13.2	9.8	16.7	0.4	0.2	1.1
83	2008-131/03:19:14	1589083001	6.43E-14	4.35	-2.49	13.2*	9.7	16.6	0.4	0.2	1

The columns of the table show the number of the spectrum (#, in order of detection); the time of detection in coordinated universal time (UTC) and as measured by the spacecraft clock (SCLK); the amplitude of charge measured at the CDA ion grid (QI (C)); the spacecraft's distance to Saturn's rotation axis (R(R_S); where $R_S = 60{,}268$ km is Saturn's radius); the spacecraft's distance to the ring plane (Z(R_S)); grain impact speeds (in kilometres per second, see below); and the corresponding grain radii estimates (in micrometres). For the grain impact speeds, the velocity V_{imp} is derived assuming that grains are in circular Keplerian orbits, and $V_{-\sigma}$ and $V_{+\sigma}$ are the $\pm 1\sigma$ of the grain impact speed distributions, inferred from an E-ring model derived from numerical simulations[73,74].
The grain radii shown in the last three columns are calculated on the basis of V_{imp}, $V_{-\sigma}$ and $V_{+\sigma}$ using equation (3.37) of ref. [78].
Events for which the UTC time is shown in bold fonts are the 64 high-quality spectra (see Methods, 'Selection of 64 high-quality spectra for Fig. 1 and Extended Data Figs. 1 and 3'). Speeds marked with an asterisk denote spectra in which one or more hydrogen cations (H^+, H_2^+, H_3^+) are present, which is used as an independent and reliable criterion for impact speeds larger than 10 km s^{-1} even if the number derived from the assumption of circular orbits (V_{imp}) is lower.

LETTER

https://doi.org/10.1038/s41586-018-0245-5

Flying couplers above spinning resonators generate irreversible refraction

Shai Maayani[1,7], Raphael Dahan[1,7], Yuri Kligerman[1], Eduard Moses[1,2], Absar U. Hassan[3], Hui Jing[4], Franco Nori[5,6], Demetrios N. Christodoulides[3] & Tal Carmon[1]*

Creating optical components that allow light to propagate in only one direction—that is, that allow non-reciprocal propagation or 'isolation' of light—is important for a range of applications. Non-reciprocal propagation of sound can be achieved simply by using mechanical components that spin[1,2]. Spinning also affects de Broglie waves[3], so a similar idea could be applied in optics. However, the extreme rotation rates that would be required, owing to light travelling much faster than sound, lead to unwanted wobbling. This wobbling makes it difficult to maintain the separation between the spinning devices and the couplers to within tolerance ranges of several nanometres, which is essential for critical coupling[4,5]. Consequently, previous applications of optical[6–17] and optomechanical[10,17–20] isolation have used alternative methods. In hard-drive technology, the magnetic read heads of a hard-disk drive fly aerodynamically above the rapidly rotating disk with nanometre precision, separated by a thin film of air with near-zero drag that acts as a lubrication layer[21]. Inspired by this, here we report the fabrication of photonic couplers (tapered fibres that couple light into the resonators) that similarly fly above spherical resonators with a separation of only a few nanometres. The resonators spin fast enough to split their counter-circulating optical modes, making the fibre coupler transparent from one side while simultaneously opaque from the other—that is, generating irreversible transmission. Our setup provides 99.6 per cent isolation of light in standard telecommunication fibres, of the type used for fibre-based quantum interconnects[22]. Unlike flat geometries, such as between a magnetic head and spinning disk, the saddle-like, convex geometry of the fibre and sphere in our setup makes it relatively easy to bring the two closer together, which could enable surface-science studies at nanometre-scale separations.

The real part of the refractive index of a material describes the speed of light in that material on the basis of linear light–matter interaction. An irreversible refractive index[23] was originally demonstrated by Fizeau in flowing water, in an experiment of paramount importance for the development of special relativity. Such an irreversible index has generally been considered in photonics only when the entire apparatus, including its detector and emitter, is moving (such as in gyroscopes). Controlling photonic structures in such a way that they exhibit an irreversible index is challenging, as is controlling rapidly moving dielectrics while self-positioning them with nanometre precision. Therefore, current photonic technology and fundamental studies that involve light generally use stationary, non-self-positioned dielectrics and rely on a reversible index that treats counter-propagating light equally. Here we report a photonic device containing a dielectric material that moves sufficiently fast, while its position is self-adjusted finely enough, to achieve an irreversible index, which either passes or blocks light depending on the direction it comes from.

We fabricate a spherical resonator by melting the end of a silica-glass cylinder, flame-polishing it and then mounting the resonator on a turbine (Fig. 1d, Methods). In our experimental setup (Fig. 1), this rotating silica sphere, which is resonant when its circumference corresponds to an integer number of optical wavelengths, is positioned near a tapered region of a standard single-mode telecommunications fibre. This tapered region is used to coupling light evanescently[4,5] into the rotating resonator. In the same manner, light is coupled out through the other side of the fibre via the same coupler. Each side of the fibre thus serves simultaneously as both an input port and an output port. Depending on the input port, light is coupled to circulate in the sphere in either the clockwise or the anticlockwise direction. Because of the Fizeau drag, the refractive index that light experiences in the spinning dielectric sphere is different in these counter-circulating directions. As a result, light entering from one side of the fibre is on resonance and is critically absorbed[4,5], whereas light entering from the other side is off resonance and so is almost 100% transmitted.

Central to the realization of such an irreversible index are the aerodynamic processes at play when the nanoscale fibre is only a small distance above the rapidly spinning spherical resonator. Numerical calculations (Methods; Fig. 2d) demonstrate that, for a sphere spinning at an angular velocity Ω, a boundary layer of air forms that is dragged past the stationary taper. By dragging the air into the region between the taper and the sphere, the moving sphere causes the taper to fly at a height h above the sphere, owing to the air pressure on the surface of the taper that faces the sphere (the 'air bearing' surface). If any perturbation causes the taper to rise higher than this stable-equilibrium height, it floats back to its original position; we refer to this behaviour as 'self-adjustment'.

As shown in Methods, the tension of the taper, T, can be estimated analytically from

$$T = 6.19\mu R^{5/2}\Omega \int_0^r (h - \sqrt{r^2 - x^2} + r)^{-3/2} \mathrm{d}x \qquad (1)$$

where μ is the viscosity of air, r is the radius of the taper and R is the radius of the sphere. When operating at tensions near 3 GPa[24], we find $h = 320$ nm and $h = 38.2$ nm for the cases shown in Figs. 2 and 3, respectively. The self-adjustment of the taper separation from the rotating sphere enables critical coupling[4,5] of light into the sphere, whereby counter-circulating light experiences optical drag identical in size, but opposite in sign.

To understand how Fizeau drag leads to isolation, we first take into account the relativistic addition of velocities when the periphery of the rotating silica resonator is moving towards or away from the input/output ports where the light coupling takes place (Fig. 1). On the basis of these considerations, we conclude that, in the laboratory frame, the refractive indices associated with the clockwise ('+') and anticlockwise ('−') whispering-gallery modes are $n_{\pm} = n[1 \pm nv(n^{-2} - 1)/c]$, where n is the refractive index of silica, c is the speed of light in vacuum, $v = R\Omega$ is the tangential velocity of the rotating optical cavity and R is the radius of the resonator.

[1]Faculty of Mechanical Engineering, Technion, Haifa, Israel. [2]J-Rom, Haifa, Israel. [3]CREOL/College of Optics and Photonics, University of Central Florida, Orlando, FL, USA. [4]Physics Department, Hunan Normal University, Changsha, China. [5]Physics Department, University of Michigan, Ann Arbor, MI, USA. [6]Theoretical Quantum Physics Laboratory, RIKEN Cluster for Pioneering Research, Wako-shi, Japan. [7]These authors contributed equally: Shai Maayani, Raphael Dahan. *e-mail: tcarmon@technion.ac.il

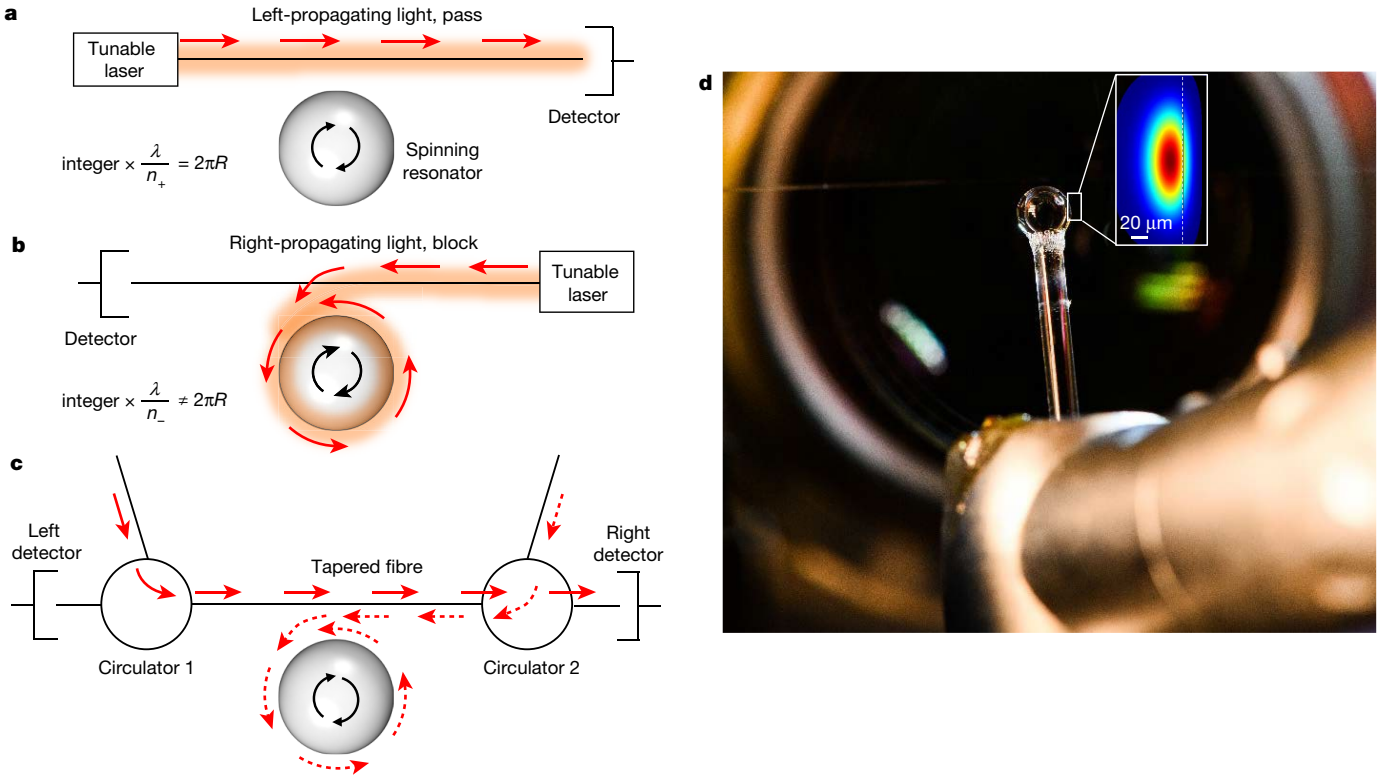


Fig. 1 | Experimental setup. Spinning the silica sphere (the resonator) results in different refractive indices for the counter-circulating whispering-gallery modes of the resonator. **a**, **b**, Illustration of the left-propagating 'pass' configuration (**a**) and the right-propagating 'block' configuration, where the input light does not match the resonance frequency of the cavity (**b**). **c**, The setup enables us to perform experiments in which light enters from both ports simultaneously. Solid (dashed) arrows represent the path of the transmitted (blocked) light. In the blocked path, the light is absorbed in the resonator and does not reach the detector. **d**, A micrograph of our experimental setup; the inset shows the calculated transverse field of the optical resonance. Optical polarization in our scheme is either parallel or perpendicular to the rotation axis of the sphere, compatible with the polarization of the transverse electric or the transverse magnetic mode of the sphere. This polarization does not affect the performance when we use our polarized light source.

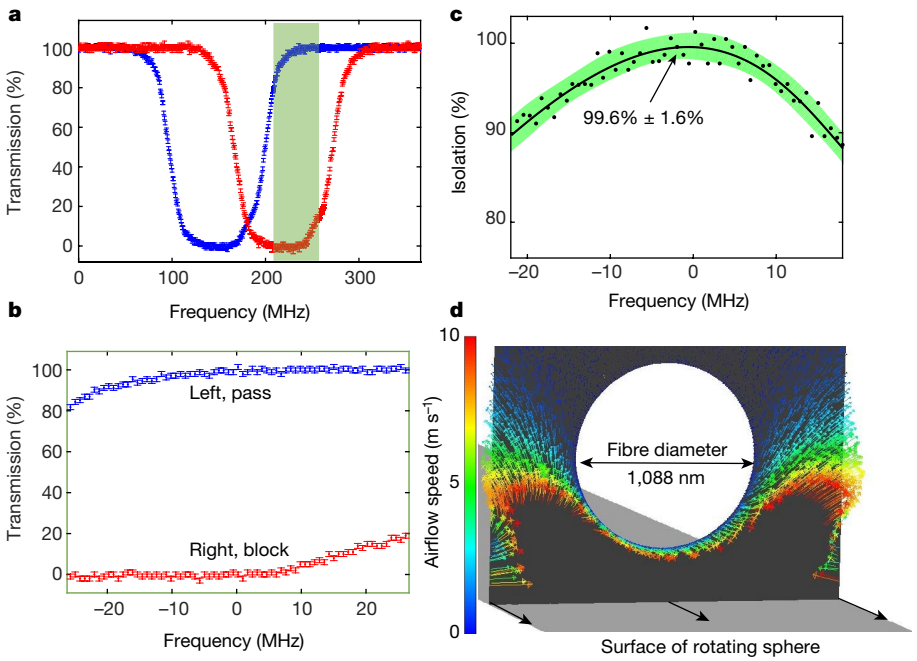


Fig. 2 | Experimentally measured isolation of 99.6%. a, **b**, Optical transmission of light coming simultaneously from the left (red) and right (blue) while scanning the laser frequency. The frequency range shown in **b** corresponds to the region shaded green in **a**, with 0 MHz corresponding to the centre of this region. The error bars show the standard deviation. **c**, Experimentally measured isolation (circles) and theoretical fit (line; see Methods); the green shading shows the confidence band for the fit. The curve peaks at a maximum isolation of 99.6% ± 1.6%. **d**, The airflow near the fibre (arrows represent calculated direction; colour represents calculated speed) pushes the taper (empty circle) away from the interface of the rotating sphere (grey surface at the bottom). For this experiment, the optical wavelength is $\lambda = 1.55\,\mu m$, the radius of the resonator is $R = 4.75$ mm and its rotation frequency is $\Omega = 2\pi \times 3{,}000$ rad s^{-1}.

The difference in resonance frequency between the counter-circulating modes is then[25]

$$\Delta\omega_{\mathrm{F}} = 2\omega_{\mathrm{rest}}\frac{nR\Omega}{c}\left(1-\frac{1}{n^2}-\frac{\lambda}{n}\frac{\mathrm{d}n}{\mathrm{d}\lambda}\right)=\eta\Omega \qquad (2)$$

where ω_{rest} is the optical resonance frequency for a stationary resonator and λ is the optical wavelength in vacuum. Scattering between counter-circulating modes[26] is eliminated in our experiments to prevent unwanted reflections (see Methods). In addition, the coupling is set to critical[4,5] to prevent transmission in the wrong direction. Under these conditions, and taking into account the frequency shift from equation (2), the rate equations[27] for the counter-circulating modes, with field amplitudes $a_{\circlearrowright}$ and $a_{\circlearrowleft}$, are

$$\begin{aligned}
&\dot{a}_{\circlearrowright}+\left[\frac{t^2}{\tau}+i(\omega-\omega_{\mathrm{rest}}+\eta\Omega)\right]a_{\circlearrowright}=i\frac{t}{\tau}\overrightarrow{a}_{\mathrm{in}}, \quad \overrightarrow{a}_{\mathrm{out}}=\left(1-\frac{t^2}{2}\right)\overrightarrow{a}_{\mathrm{in}}+ita_{\circlearrowright}\\
&\dot{a}_{\circlearrowleft}+\left[\frac{t^2}{\tau}+i(\omega-\omega_{\mathrm{rest}}-\eta\Omega)\right]a_{\circlearrowleft}=i\frac{t}{\tau}\overleftarrow{a}_{\mathrm{in}}, \quad \overleftarrow{a}_{\mathrm{out}}=\left(1-\frac{t^2}{2}\right)\overleftarrow{a}_{\mathrm{in}}+ita_{\circlearrowleft}
\end{aligned} \qquad (3)$$

where ω is the optical frequency of the input light, t is the real amplitude coefficient of transmittance between the fibre and the resonator, τ is the circulation time for the mode travelling inside the sphere, $\overrightarrow{a}_{\mathrm{in}}$ ($\overleftarrow{a}_{\mathrm{in}}$) is the optical input from the left (right) and $\overrightarrow{a}_{\mathrm{out}}$ ($\overleftarrow{a}_{\mathrm{out}}$) is the optical output from the right (left). Accordingly, the optical transmission for the opposing directions are

$$\begin{aligned}
\overrightarrow{T}&=\left|\frac{\overrightarrow{a}_{\mathrm{out}}}{\overrightarrow{a}_{\mathrm{in}}}\right|^2=1-\frac{t^4}{t^4+\tau^2(\omega-\omega_{\mathrm{rest}}+\eta\Omega)^2}\\
\overleftarrow{T}&=\left|\frac{\overleftarrow{a}_{\mathrm{out}}}{\overleftarrow{a}_{\mathrm{in}}}\right|^2=1-\frac{t^4}{t^4+\tau^2(\omega-\omega_{\mathrm{rest}}-\eta\Omega)^2}
\end{aligned} \qquad (4)$$

Equation (4) clearly indicates that the transmission in this rotating resonator system is non-reciprocal: the equations are identical, except for the sign of η, which facilitates isolation by resonantly absorbing light entering from one side, while off-resonantly transmitting the counter-propagating light. Importantly, the equations in equation (3) are uncoupled because $\overrightarrow{a}_{\mathrm{out}}$ is not a function of $\overleftarrow{a}_{\mathrm{in}}$ and $\overleftarrow{a}_{\mathrm{out}}$ is not a function of $\overrightarrow{a}_{\mathrm{in}}$, which allows crosstalk-free operation while light enters simultaneously from both sides.

We measured the spectral transmission as a function of the input direction (Fig. 2) and as a function of rotation speed (Fig. 3). In the first experiment, we demonstrate that our isolator almost completely transmits light coming from one side of the fibre while blocking light coming from the other side. As expected from equation (4), tuning the laser frequency ω through the counter-circulating resonance frequencies, $\omega_{\mathrm{rest}}-\eta\Omega$ and $\omega_{\mathrm{rest}}+\eta\Omega$, reveals oppositely drifted dips in transmission in for the two opposing directions (Fig. 2a). The split between these counter-circulating modes is larger than their widths, which enables light from one side to pass through while the other side is blocked (Fig. 2b). Isolation is most challenging when light enters from both sides of the isolator at the same time, because the isolator might suffer from both imperfect transmission of one beam and backscattering of the other.

To measure the degree of isolation, we subtract the power of light in the blocking direction (ideally 0%; red, Fig. 2b) from the power of light in the transmitting direction (ideally 100%; blue, Fig. 2b), while injecting light to the isolator from both sides (Fig. 1c). As can be seen in Fig. 2c, we measure a remarkable 99.6% intra-fibre isolation (Methods), despite the challenge of simultaneous two-port input.

In the second experiment, we measure the frequency shift as a function of the mechanical rotation rate (Fig. 3). Starting from rest ($\Omega=0$) as a control experiment, the counter-circulating modes overlap, owing to their expected degeneracy (Fig. 3b). As predicted by equation (2),

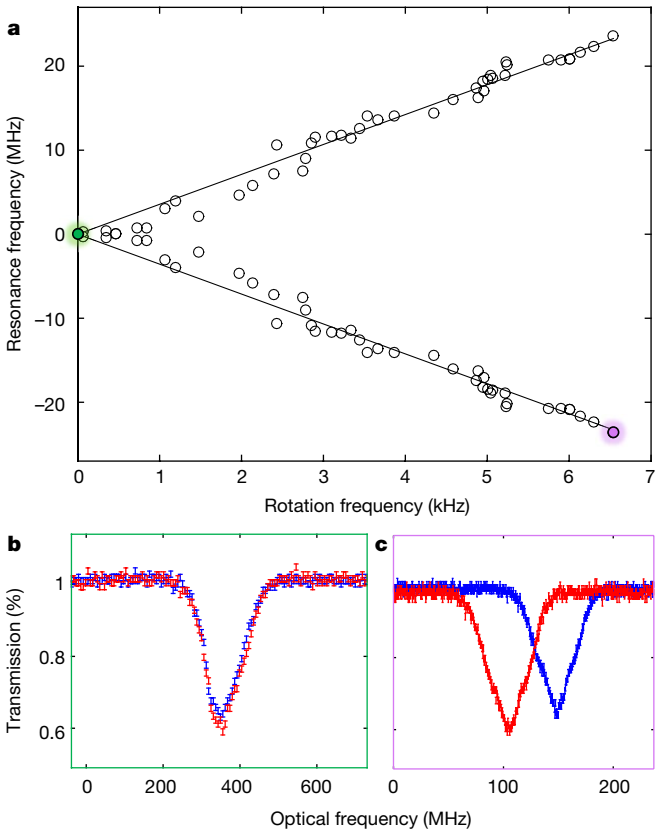


Fig. 3 | The Fizeau shift. a, The Fizeau shift is evident from the split between the counter-circulating resonances as a function of the rotation speed of the resonator. Circles describe exprimental results and solid lines represent linear fits. **b**, **c**, Transmission spectrum in a control experiment while the resonator is at rest (**b**; corresponding to the green shaded point in **a**) and while the resonator is spinning at 6.6 kHz (**c**; purple shaded point in **a**). The experimental setup is shown in Fig. 1. The radius of the sphere is $R = 1.1$ mm. Red (blue) data points are for transmission of light coming from the right (left). The error bars are one standard deviation.

increasing the mechanical rotation frequency Ω results in a linear opposing frequency shift of $\eta\Omega$ (Fig. 3a) for the counter-circulating modes.

In conclusion, we have demonstrated optical non-reciprocity experimentally by breaking time-reversal symmetry through a mechanically spinning optical resonator. In addition, enabling intra-fibre isolation could be beneficial for applications such as fibre-linked quantum photon routings[22]. We believe that our work can be extended to spinning microphotonics, with similar nanometre-scale flying heights. When approaching couple–resonator separations of 300 fm[28] (a regime related to femtotechnology), repulsive van der Waals forces are predicted to start acting against the large Casimir attraction[28]. Being highly convex, spheres might be ideal for testing exceptionally small separation distances, near the boundary between the compliant surfaces. Furthermore, with the availability of quadrant photodiodes for highly accurate measurements in optical tweezers, tweezed counter-rotating spheres, one flying the other, might allow tests of gravity at distances shorter than what the current state-of-the-art permits. Although challenging, miniaturization and faster spinning of our resonators might be possible with continual technological improvements in the rotation[29] and manipulation[30] of spherical dielectrics.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0245-5.

Received: 2 November 2017; Accepted: 17 April 2018;
Published online 27 June 2018.

1. Fleury, R., Sounas, D. L., Sieck, C. F., Haberman, M. R. & Alù, A. Sound isolation and giant linear nonreciprocity in a compact acoustic circulator. *Science* **343**, 516–519 (2014).
2. Yang, Z. et al. Topological acoustics. *Phys. Rev. Lett.* **114**, 114301 (2015).
3. Hasselbach, F. & Nicklaus, M. Sagnac experiment with electrons: observation of the rotational phase shift of electron waves in vacuum. *Phys. Rev. A* **48**, 143–151 (1993).
4. Dubreuil, N. et al. Eroded monomode optical fiber for whispering-gallery mode excitation in fused-silica microspheres. *Opt. Lett.* **20**, 813–815 (1995).
5. Spillane, S. M., Kippenberg, T. J., Painter, O. J. & Vahala, K. J. Ideality in a fiber-taper-coupled microresonator system for application to cavity quantum electrodynamics. *Phys. Rev. Lett.* **91**, 043902 (2003).
6. Chang, L. et al. Parity-time symmetry and variable optical isolation in active-passive-coupled microresonators. *Nat. Photon.* **8**, 524–529 (2014).
7. Gallo, K., Assanto, G., Parameswaran, K. R. & Fejer, M. M. All-optical diode in a periodically poled lithium niobate waveguide. *Appl. Phys. Lett.* **79**, 314–316 (2001).
8. Ibrahim, S. K., Bhandare, S., Sandel, D., Zhang, H. & Noe, R. Non-magnetic 30 dB integrated optical isolator in III/V material. *Electron. Lett.* **40**, 1293–1294 (2004).
9. Yu, Z. & Fan, S. Complete optical isolation created by indirect interband photonic transitions. *Nat. Photon.* **3**, 91–94 (2009); corrigendum 3, 303 (2009).
10. Kang, M. S., Butsch, A. & Russell, P. S. J. Reconfigurable light-driven opto-acoustic isolators in photonic crystal fibre. *Nat. Photon.* **5**, 549–553 (2011).
11. Lira, H., Yu, Z., Fan, S. & Lipson, M. Electrically driven nonreciprocity induced by interband photonic transition on a silicon chip. *Phys. Rev. Lett.* **109**, 033901 (2012).
12. Fan, L. et al. An all-silicon passive optical diode. *Science* **335**, 447–450 (2012).
13. Poulton, C. G. et al. Design for broadband on-chip isolator using stimulated Brillouin scattering in dispersion-engineered chalcogenide waveguides. *Opt. Express* **20**, 21235–21246 (2012).
14. Hafezi, M. & Rabl, P. Optomechanically induced non-reciprocity in microring resonators. *Opt. Express* **20**, 7672–7684 (2012).
15. Peng, B. et al. Parity-time-symmetric whispering-gallery microcavities. *Nat. Phys.* **10**, 394–398 (2014).
16. Lu, L., Joannopoulos, J. D. & Soljačić, M. Topological photonics. *Nat. Photon.* **8**, 821–829 (2014).
17. Shen, Z. et al. Experimental realization of optomechanically induced non-reciprocity. *Nat. Photon.* **10**, 657–661 (2016).
18. Kim, J., Kim, S. & Bahl, G. Complete linear optical isolation at the microscale with ultralow loss. *Sci. Rep.* **7**, 1647 (2017).
19. Fang, K. et al. Generalized non-reciprocity in an optomechanical circuit via synthetic magnetism and reservoir engineering. *Nat. Phys.* **13**, 465–471 (2017).
20. Ruesink, F., Miri, M.-A., Alù, A. & Verhagen, E. Nonreciprocity and magnetic-free isolation based on optomechanical interactions. *Nat. Commun.* **7**, 13662 (2016).
21. Gross, W. A. *Gas Film Lubrication* (Wiley, New York, 1962).
22. Shomroni, I., Rosenblum, S., Lovsky, Y. & Bechler, O. All-optical routing of single photons by a one-atom switch controlled by a single photon. *Science* **345**, 903–906 (2014).
23. Franke-Arnold, S., Gibson, G., Boyd, R. W. & Padgett, M. J. Rotary photon drag enhanced by a slow-light medium. *Science* **333**, 65–67 (2011).
24. Matthewson, M., Kurkjian, C. R. & Gulati, S. T. Strength measurement of optical fibers by bending. *J. Am. Ceram. Soc.* **69**, 815–821 (1986).
25. Malykin, G. B. The Sagnac effect: correct and incorrect explanations. *Phys. Uspekhi* **43**, 1229–1252 (2000).
26. Mazzei, A. et al. Controlled coupling of counterpropagating whispering-gallery modes by a single Rayleigh scatterer: a classical problem in a quantum optical light. *Phys. Rev. Lett.* **99**, 173603 (2007).
27. Gorodetsky, M. L. & Ilchenko, V. S. Optical microsphere resonators: optimal coupling to high-*Q* whispering-gallery modes. *J. Opt. Soc. Am. B* **16**, 147–154 (1999).
28. Li, J., Liu, B., Hua, W. & Ma, Y. Effects of intermolecular forces on deep sub-10 nm spaced sliders. *IEEE Trans. Magn.* **38**, 2141–2143 (2002).
29. Arita, Y., Mazilu, M. & Dholakia, K. Laser-induced rotation and cooling of a trapped microgyroscope in vacuum. *Nat. Commun.* **4**, 2374 (2013).
30. Grier, D. G. A revolution in optical manipulation. *Nature* **424**, 810–816 (2003).

Acknowledgements We thank U. Hofi, Z. Katz, Y. Halupovich and B. Khachatryan for their help. This work was funded by the Israeli Centers for Research Excellence (I-CORE), 'Circle of Light' Excellence Center, the Israel Science Foundation (2013/15), the Israel Ministry of Science, Technology and Space, the MURI Center for Dynamic Magneto-Optics via the AFOSR Award number FA9550-14-1-0040, the Army Research Office (ARO) under grant number 73315PH, the AOARD under grant number FA2386-18-1-4045, the CREST under grant number JPMJCR1676, the IMPACT programme of JST, the RIKEN-AIST Challenge Research Fund, the JSPS-RFBR under grant number 17-52-50023, and the Sir John Templeton Foundation.

Reviewer information *Nature* thanks A. Alù and M. Levy for their contribution to the peer review of this work.

Author contributions S.M. and R.D. performed the experiments. A.U.H., H.J., F.N., E.M., Y.K. and D.N.C. performed the theoretical analysis. T.C. supervised the work.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0245-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Backscattering and bandwidth. Our resonators have a relatively low optical quality factor ($Q < 10^6$), which results in a larger operation bandwidth and reduces unwanted backscattering.

Bandwidth. Low-Q resonators have broader resonance full-width at half-maxima and therefore improved bandwidth.

Backscattering. Backscattering typically appears in resonators with $Q \approx 10^8$, where absorption is reduced but scattering is still large. In such ultrahigh-Q resonators, frozen thermal capillary waves on the resonator solid-phase boundaries are created during reflow and scatter light later. Because our quality factor is relatively low ($Q < 10^6$) and probably governed by absorption rather than scattering, we did not observe any traces of backscattering in the spectrum of our rotating-resonator isolator (Fig. 2a, b). On the basis of the rotating-resonator isolation ratio (Fig. 2c), we conclude that imperfections in isolator performance, including those resulting from potential backscattering, are relevant to less than 0.2% of the input power. Although we do not exclude the possibility of backscattering effects in rotating-resonator isolators, in particular when the surface is rough, we did not observe such backscattering in our experiment.

Fabrication. We rotate a silica cylinder while its end is heated with a hydrogen–oxygen flame to reflow into a spherical shape. We did not attempt to increase the optical quality factor because it could have reduced the operation bandwidth of our isolator. For this reason, we use industrial-grade (rather than ultra-pure) gases for our flame.

Measuring and calculating isolation. Here we define isolation as the left-propagating (pass) transmission less the right-propagating (block) transmission. For the experimental results (black circles, Fig. 2c), the transmissions subtracted are those measured experimentally (Fig. 2a). For the theoretical fit (solid line, Fig. 2c), the subtracted transmissions are those revealed from fitting the experimental results (Fig. 2a) to equation (4).

The flying height. We observe experimentally that the taper does not touch the rotating resonator even if pushed towards it. Indications that the taper veers away from the spinning resonator when pushed towards it include our experimental observation that there is no wear of the taper or sphere, even after long periods of operation, and the fact that the taper does not stick to the rotating resonator, even when pushed towards it. By contrast, in a control experiment in which the sphere is stationary, the taper does stick to the sphere, through van der Waals forces, and needs to be pulled back to break the connection. Using a microscope, we then see micrometre-scale craters on the surfaces of the taper and the sphere, where contact was established. No such adhesion or abrasion is observed for the rotating-resonator case.

Fluid-film lubrication[31] between surfaces provides a simple way to achieve the desired motion of a machine element with minimal friction and no wear. Such bearings have been studied for almost a century and found to be robust. In more detail, liquid- and gas-film[21] lubrication have been studied for applications such as sand-proof bearings in marine vessels, dating back to 1932[32], and have enabled flying heights of less than 10 nm in recently demonstrated helium-filled hard drives. Although such fliers are generally stable, modelling of them should consider effects such as lubricant compressibility and intermolecular forces.

Reynolds equation has previously been solved[33] with boundary conditions of a foil wrapping a rotating cylinder. Using similar methods, but with some modification[31], the gap between the surface of the rotating cylinder and the foil was calculated to be

$$h = 0.643R^{5/3}\left(\frac{6\mu\Omega}{t'}\right)^{2/3} \tag{5}$$

where t' is the tension of the foil per unit width, R and Ω are the radius and angular velocity of the cylinder and μ is the viscosity of the fluid. In our case, the curvature of the sphere along the polar direction is so small compared with the width of the film-lubricated region of interest that it thins by only 1 Å by the end of this region. For this reason, we can estimate our rotating sphere using equation (5), even though it was originally developed for a rotating cylinder. Because equation (5) deals with flat foil, whereas our flyer is a wire, we break the circular cross-section of our wire into a set of infinitesimal flat stairs. We integrate the tension over these stairs to get the total tension T as a function of the minimum taper–sphere separation h (see equation (1)). Under typical operation at a fibre tensile stress near 3 GPa[24], equation (1) (with some corrections explained below) yields $h = 38.2$ nm for the sphere with $R = 1.1$ mm (Fig. 3) and $h = 320$ nm for the sphere with $R = 4.75$ mm (Fig. 2).

We now check the various assumptions of equation (1).

Intermolecular forces. It has been shown that effects such as Casimir and van der Waals forces begin to attract the flyer towards the rotor when the gap between them is reduced to less than 10 nm, and to strongly repel them when the gap is narrowed further, normally to below 300 fm[28]. Taking intermolecular forces into account[28], equation (1) becomes

$$T = 6.19\mu R^{5/2}\Omega\int_0^r\left(\frac{1}{h-\sqrt{r^2-x^2}+r}\right)^{3/2}\mathrm{d}x + rR\left(-\frac{A}{6\pi h^3}+\frac{B}{45\pi h^9}-\frac{\pi^2 c\hbar}{240h^4}\right) \tag{6}$$

where A and B are the Hamaker constants and $\hbar$ is the reduced Planck constant. As expected, this intermolecular-force correction was found to be negligible in our experiment, with a contribution to the tension of less than 1% of the gas-lubricant contribution. Nonetheless, we do not exclude the possibility that such intermolecular forces might be relevant in future experiments, for example, when the rotation speed is reduced and the optical flyer is closer to the dielectric sphere.

Lubricant compressibility. We considered the effects of lubricant compressibility, using previously described methods[34], and found them to be relevant. In more detail, air was calculated to be up to 6.6 times denser at the lubrication region, requiring a correction of the constant in equation (1)[31] (see Extended Data Table 1).

Tapered-fibre stiffness. We calculated the effects of the tapered-fibre stiffness, using a previously reported method[35], and found that they satisfy a condition that demonstrates the validity of equation (1).

Wrap angle. We calculated the effects of wrap angle, using a previously reported method[36], and found that they introduce a correction of less than 1% in the constant coefficient in equation (1).

Fluid inertia. We calculated the effects of fluid inertia were calculated, using a previously reported method[37], and found that they satisfy a condition that demonstrates the validity of equation (1).

Mean free pass. At a certain small scale, we can no longer consider air as a continuum and so have to consider the mean free path[37] between individual air molecules. This issue in the gas-film lubricant is treated here by using the Knudsen number, defined by the ratio between the flying height h and the mean free pass[31].

For the case shown in Fig. 2, the Knudsen number is 0.13—corresponding to a regime in which continuous flow can be assumed and equation (1) needs no corrections.

For the case shown in Fig. 3, the Knudsen number is 1.4—corresponding to a regime in which free molecular flow effects should be taken into account. Calculating such flows requires a molecular-based model such as the Fukui–Kaneko model[38]. Several other models for molecular gas-film lubrication have been reviewed[39] to provide an order-of-magnitude variation in results at this Knudsen number.

Numerically calculating the field flow. We calculate the airflow near the taper numerically using Ansys–Fluent software (https://www.ansys.com), assuming compressible flow.

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

31. Gross, W. A. et al. *Fluid Film Lubrication* (John Wiley and Sons, New York, 1980).
32. Busse, W. F. & Denton, W. H. Water-lubricated soft-rubber bearings. *Trans. Am. Soc. Mech. Eng.* **54**, 3–10 (1932).
33. Blok, H. & Van Rossum, J. J. The foil bearing—a new departure in hydrodynamic lubrication. *Lubr. Eng.* **9**, 316–320 (1953).
34. Eshel, A. Compressibility effects on the infinitely wide, perfectly flexible foil bearing. *J. Lubr. Technol.* **90**, 221–225 (1968).
35. Eshel, A. & Elrod, H. G. Stiffness effects on the infinitely wide foil bearing. *J. Lubr. Technol.* **89**, 92–97 (1967).
36. Langlois, W. E. The lightly loaded foil bearing at zero angle of wrap. *IBM J. Res. Develop.* **7**, 112–116 (1963).
37. Jennings, S. G. The mean free path in air. *J. Aerosol Sci.* **19**, 159–166 (1988).
38. Fukui, S. & Kaneko, R. Analysis of ultra-thin gas film lubrication based on linearized Boltzmann equation: first report—derivation of a generalized lubrication equation including thermal creep flow. *J. Tribol.* **110**, 253–261 (1988).
39. Shen, S. & Chen, G. in *Encyclopedia of Tribology* (eds Wang, Q. J. & Chung, Y.-W.) 2309–2313 (Springer, New York, 2013).

Extended Data Table 1 | Effects of lubricant compressibility

Case	Lubricant compressibility [-]	Compressibility reduces the constant in equation (1) by	Flying height [nm]
Figure 2	6.6	56%	320
Figure 3	1.4	28%	38.2

LETTER

https://doi.org/10.1038/s41586-018-0223-y

A standing molecule as a single-electron field emitter

Taner Esat[1,2], Niklas Friedrich[1,2], F. Stefan Tautz[1,2]* & Ruslan Temirov[1,2]

Scanning probe microscopy makes it possible to image and spectroscopically characterize nanoscale objects, and to manipulate[1–3] and excite[4–8] them; even time-resolved experiments are now routinely achieved[9,10]. This combination of capabilities has enabled proof-of-principle demonstrations of nanoscale devices, including logic operations based on molecular cascades[11], a single-atom transistor[12], a single-atom magnetic memory cell[13] and a kilobyte atomic memory[14]. However, a key challenge is fabricating device structures that can overcome their attraction to the underlying surface and thus protrude from the two-dimensional flatlands of the surface. Here we demonstrate the fabrication of such a structure: we use the tip of a scanning probe microscope to lift a large planar aromatic molecule (3,4,9,10-perylenetetracarboxylic-dianhydride) into an upright, standing geometry on a pedestal of two metal (silver) adatoms. This atypical and surprisingly stable upright orientation of the single molecule, which under all known circumstances adsorbs flat on metals[15,16], enables the system to function as a coherent single-electron field emitter. We anticipate that other metastable adsorbate configurations might also be accessible, thereby opening up the third dimension for the design of functional nanostructures on surfaces.

The fabrication of a standing molecule on an Ag(111) surface (Fig. 1a) proceeds in three steps. First, two silver adatoms are attached

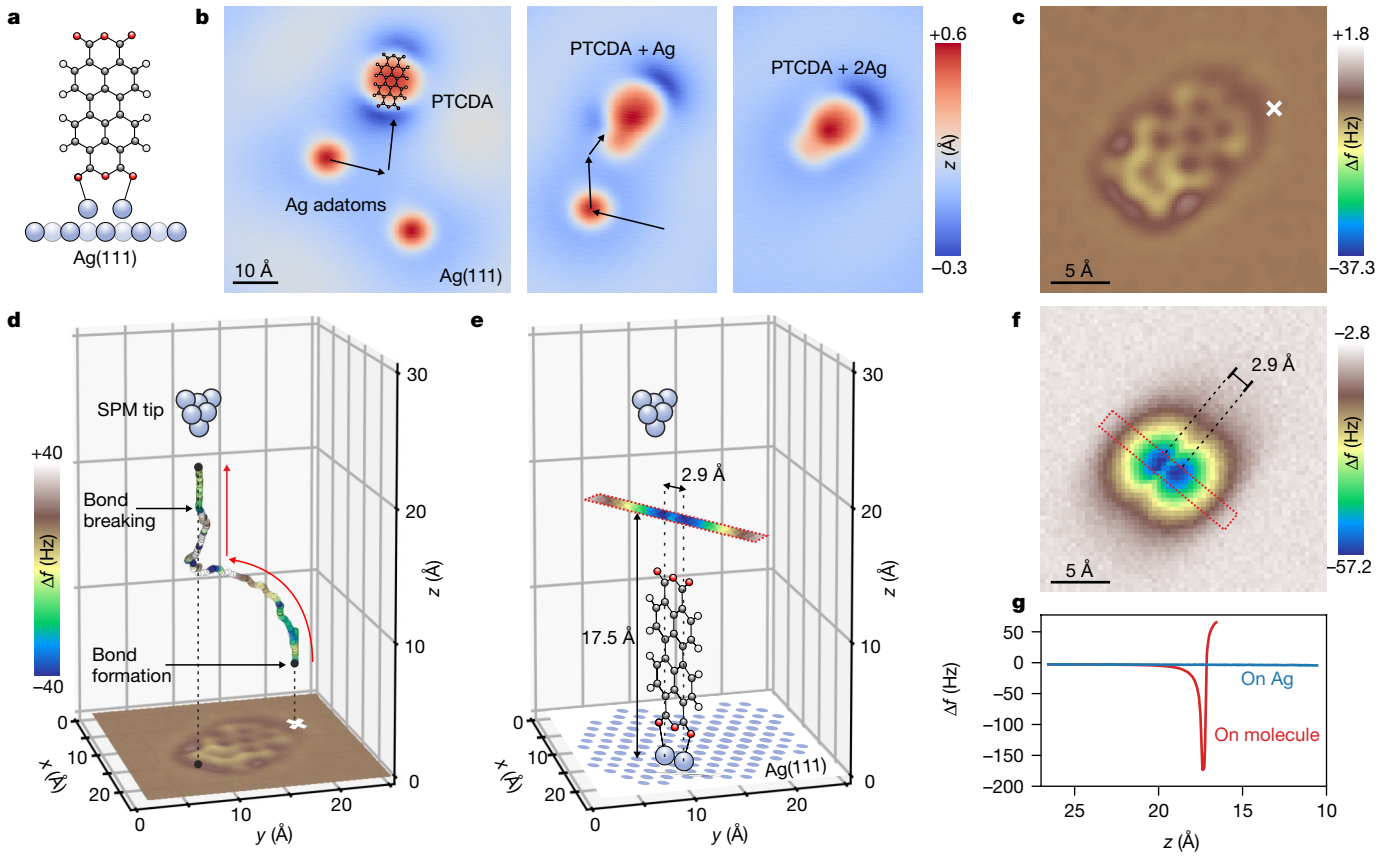


Fig. 1 | Creation of a standing molecule. a, Schematic side view of a standing PTCDA molecule. **b**, Constant-current STM images show the stepwise assembly of a PTCDA + 2Ag complex by lateral manipulation of silver atoms with the tip of the SPM. The structure of PTCDA has been overlaid. **c**, AFM image of the PTCDA + 2Ag complex, measured with a tip that has been functionalized with a carbon monoxide molecule (see Methods). **d**, The three-dimensional trajectory that was used to lift up the PTCDA + 2Ag complex. The trajectory proceeds from the contact point (labelled 'bond formation' and located above the white cross in **c** and **d**) in two segments (indicated by red arrows), through the bond-breaking point. The colour of the trajectory refers to the measured AFM signal Δf. **e**, Side view of the standing molecule, including the silver atoms at the surface contact. The bar shows the AFM signal Δf at tip height of $z = 17.5$ Å above the surface (region in **f** indicated by the red dotted box). **f**, AFM image of the standing molecule, recorded at $z = 17.5$ Å. **g**, Approach curves of the SPM tip above the standing molecule (red) and above the bare Ag(111) surface (blue). The ordinate shows the frequency shift Δf of the AFM, which is proportional to the vertical force gradient.

[1]Peter Grünberg Institute (PGI-3), Forschungszentrum Jülich, Jülich, Germany. [2]Jülich Aachen Research Alliance (JARA), Fundamentals of Future Information Technology, Jülich, Germany. *e-mail: s.tautz@fz-juelich.de

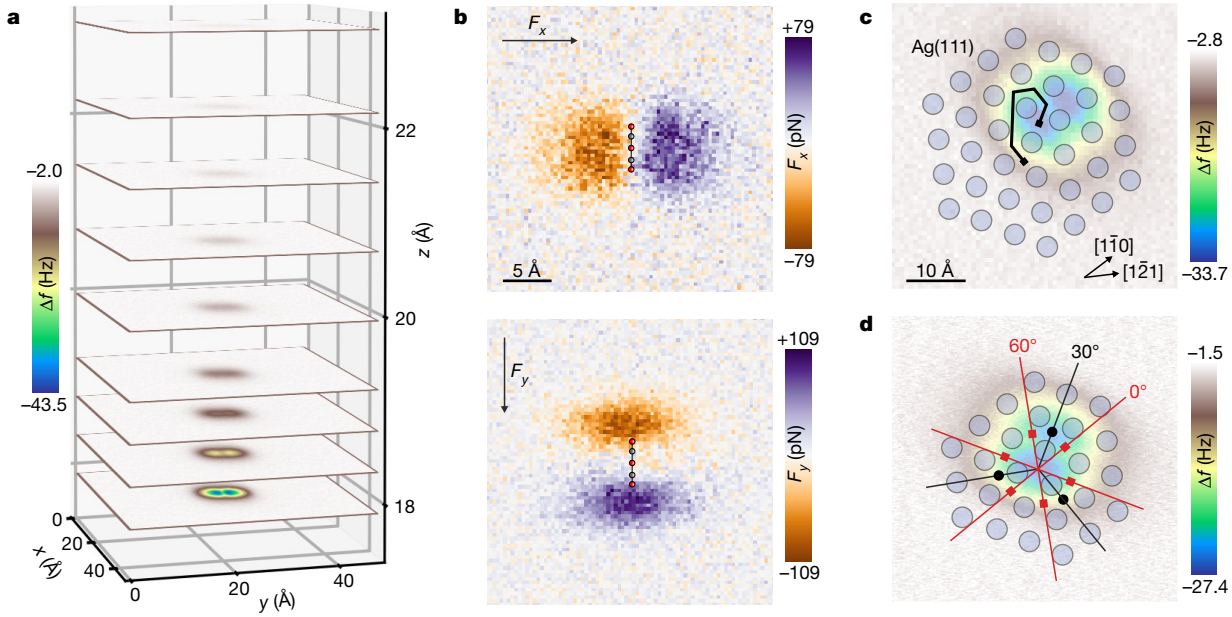


Fig. 2 | Stability and geometry of the standing molecule. a, AFM images above a standing molecule. For clarity, only selected images of the dense stack, measured in steps of $\Delta z = 0.15$ Å, are shown. **b**, Lateral forces F_x (top) and F_y (bottom) exerted by the tip on the standing molecule at $z = 17.5$ Å. **c**, Lateral manipulation trajectory (black line). Only hollow sites of the Ag(111) surface are visited. **d**, Orientation of the standing molecule, as derived from Extended Data Fig. 5. Black and red symbols indicate the possible positions of one of the silver atoms at the surface contact, when the other sits in the centre. During both translations and rotations, the standing molecule never jumps over a surface atom. Images in **c** and **d** were recorded at $z = 17.5$ Å. The lateral and rotational manipulation procedures are shown in Extended Data Fig. 3.

to one of the anhydride groups of the 3,4,9,10-perylenetetracarboxylic-dianhydride (PTCDA) molecule that we use in this study[17,18]. This proceeds by in-plane manipulation with the tip of a scanning probe microscope (SPM)[2], as shown in the series of scanning tunnelling microscopy (STM) images in Fig. 1b. The high-resolution atomic force microscopy (AFM) image in Fig. 1c reveals an asymmetry of the PTCDA + 2Ag complex, which demonstrates that the chemical structure of PTCDA has been modified. It also shows that the complex sits on the surface in an essentially planar configuration, with a tilt of 4.0° ± 0.6°, as determined using the method reported in ref. [19].

In the second step, the tip is moved towards the carboxylic oxygen on the pristine side of the molecule (cross in Fig. 1c). During approach, the current I through the junction is monitored. The approach is stopped immediately after the current increases abruptly, because this indicates the formation of a covalent bond between the tip and the molecule[20].

In the third step, the tip with the molecule attached is retracted along a customized three-dimensional trajectory (Fig. 1d)[21]. Initially, this trajectory mimics the peeling of an adhesive from a desk, carried out with the constraint that the far end of the molecule (where the silver atoms are attached) remains stationary on the surface. Later, the tip is lifted vertically, straining the tip–molecule–surface junction of the now-upright molecule until the bond between the molecule and the tip breaks. It is clear that this bond is weaker than the bond to the surface, because the latter consists of two Ag–O bonds, whereas the former has only one.

The AFM image in Fig. 1f, scanned 17.5 Å above the surface, reveals a prolate object with linear dimensions of 7.8 Å × 5.9 Å × 17.5 Å (Fig. 1f, g), close to what is expected for an upright molecule. This demonstrates that after the bond to the tip has broken the molecule remains standing on the surface (Fig. 1e). Similar molecules to PTCDA can also be erected (Extended Data Fig. 1), and two standing molecules can be placed at desired positions relative to each other (Extended Data Fig. 2).

In our experience, the standing molecule never falls back to the surface by itself. To learn more about this remarkable stability of the standing molecule, we study its interaction with the tip. We acquire AFM images at different heights above the molecule (Fig. 2a) and integrate the frequency-shift signal $\Delta f(x, y, z)$, which is proportional to the force gradient $\partial F_z(x, y, z)/\partial z$, twice along z at each lateral position (x, y) (ref. [2]). We thus obtain the full interaction potential $U(x, y, z)$ between the tip and the standing molecule. Figure 2b displays (x, y) maps of the lateral forces $F_x = -\partial U/\partial x$ and $F_y = -\partial U/\partial y$ that act on the molecule for a tip height just above the upper end of the molecule. These forces are attractive, reaching maximum values of $F_x \approx 80$ pN and $F_y \approx 110$ pN. For larger forces, which act at even closer distances, the molecule jumps laterally on the surface towards the tip, keeping its upright orientation (Extended Data Fig. 3a). We therefore did not succeed in toppling the molecule over mechanically. Note that the threshold force for the lateral displacement of the standing molecule is similar to the forces required to move single atoms[2].

The force constant for tilting the molecule out of the upright orientation can be estimated from Fig. 2b (see Methods and Extended Data Fig. 4 for details). The standing molecule can also be rotated around its long axis by injecting a current (Extended Data Fig. 3b). By combining translational and rotational manipulation, we can determine the structure at the interface between the standing molecule and the Ag(111) surface. We find that the standing molecule adsorbs with its silver atoms in hollow sites of the Ag(111) surface (Fig. 2c). The epitaxial registry between the essentially rigid Ag–O arrangement in the standing molecule and the hollow sites in the Ag(111) surface determines the possible orientations of the standing molecule (Fig. 2d, Extended Data Fig. 5).

We now discuss the charging and field-emission properties of the standing molecule. The corresponding processes are shown schematically in Fig. 3a. In the standing configuration, there is negligible overlap between the metal surface and the lowest unoccupied molecular orbital (LUMO) of the standing molecule—a π orbital that has its lobes oriented perpendicular to the plane of the molecule. Hence, the electronic coupling between this orbital and the surface is weak, and the charge state of the molecule, functioning as a quantum dot, can be controlled with single-electron precision. When a negative bias voltage V is applied to the surface, the molecular levels are gated with respect to the Fermi energy E_F of the substrate. At a critical bias voltage, the unoccupied molecular orbital shown in Fig. 3a aligns with E_F and one electron tunnels from the surface into this orbital. This charging event appears in the Δf signal of the AFM as a sharp dip[3,22] (red curve in Fig. 3a) and in the field-emission current as a steep rise (blue curve). Once on

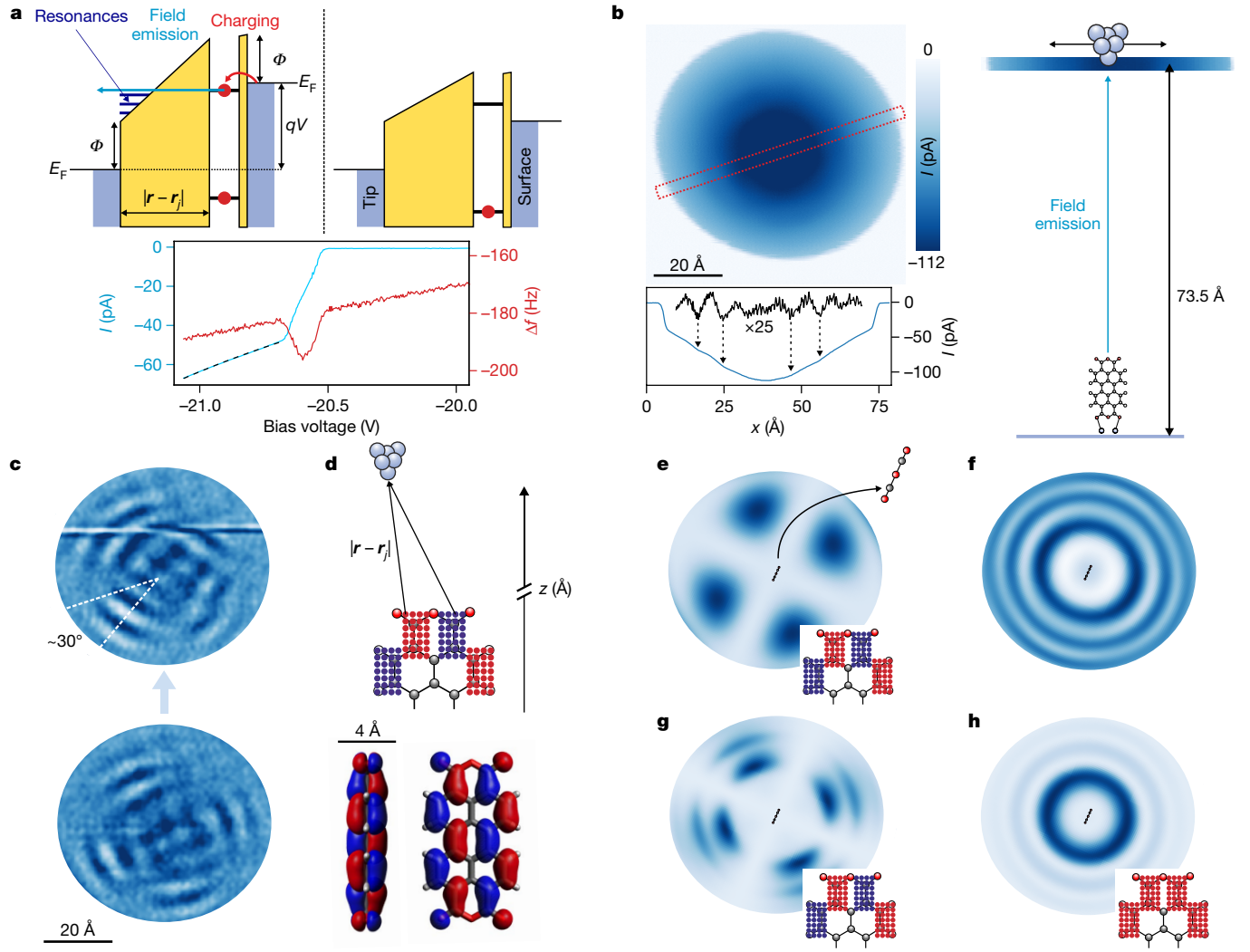


Fig. 3 | A standing molecule as a single-electron field emitter. a, Energy-level diagrams before charging (top right) and in the field-emission regime (top left), along with the frequency shift $\Delta f(V)$ (red) and field-emission current $I(V)$ (blue; the dotted black line is a Fowler–Nordheim fit) at $z = 56.0$ Å (bottom). E_F, Fermi energy; q, elementary charge; V, bias voltage; Φ, work function; $|\boldsymbol{r}-\boldsymbol{r}_j|$, distances between the tip and the field-emission points. **b**, Field-emission image (top left) and line profile (bottom left; blue) above a standing molecule at $z = 73.5$ Å, with $V = -24.7$ V. The black line shows a profile after filtering out the background, and multiplied by a factor of 25. The signal-to-background ratio is about 2%. A scale drawing of the field-emission experiment is shown on the right. **c**, Successive field-emission images (without the background), taken at the same conditions as for **b**. The arrow indicates the time line of the experiment. The two dotted white lines indicate the rotation of the interference pattern, and hence of the molecule, by 30°, from a 30° orientation to a 60° orientation, as defined in Fig. 2d. **d**, Bottom, side and top view of the LUMO of PTCDA[30]. The colour indicates the phase of the wavefunction: $\varphi_j = 0$ (red) or $\varphi_j = \pi$ (blue). Top, interference of spherical waves from different points $\boldsymbol{r}_j$ within an extended orbital, at the position $\boldsymbol{r}$ of the tip. The lobes of the LUMO of PTCDA are approximated by grid points located within cuboids. **e**, Interference pattern $|\sum_j \Psi_j(\boldsymbol{r})|^2$. **f**, Gundlach oscillations $\left|\sum_j T_j(V, \boldsymbol{r})\right|^2$. **g**, Calculated field-emission pattern $|\sum_j T_j(V, \boldsymbol{r})\Psi_j(\boldsymbol{r})|^2$. Images in **e**–**g** were calculated using grid and phases φ_j as in **d** (see insets), and with z and V as in **b** and **c**. **h**, As for **g**, but with $\varphi_j = 0$ everywhere. All field-emission images are shown in the orientation of Fig. 2c, d.

the standing molecule, the additional electron may tunnel through the trapezoidal barrier to the tip (Fig. 3a), giving rise to a field-emission current that follows the well-known Fowler–Nordheim behaviour $I(V) \propto (V^2/\Phi)\exp(-\beta|\boldsymbol{r} - \boldsymbol{r}_j|\Phi^{3/2}/|V|)$, where Φ is the work function and β is a constant[23]. For the fit in Fig. 3a, we have replaced $|\boldsymbol{r} - \boldsymbol{r}_j|$ by the vertical distance d between the top of the standing molecule and the tip.

An image of the field-emission current in the detection plane above the standing molecule is shown in Fig. 3b. Outside an elliptical area there is no field emission. This result is a consequence of the fact that for lateral tip positions too far away from the standing molecule the gating is too weak to provoke electron transfer into the unoccupied level in Fig. 3a, which also precludes field emission. (It also shows that in our experiment field emission occurs exclusively through the standing molecule.) The current appears to increase exponentially towards the centre of the elliptical field-emission area (Fig. 3b), in line with the Fowler–Nordheim equation. However, closer inspection reveals additional structure. After filtering out the background that originates from field-emitted electrons that hit the tip away from its apex, we find a pattern of concentric rings that are modulated by a four-fold-symmetric cloverleaf (Fig. 3c). Figure 3c also shows that the pattern is locked to the molecule: after about one-third of the second of two consecutive images, the molecule and with it the field-emission pattern rotated by about 30°.

Modelling the pattern in Fig. 3c demonstrates that the standing molecule is a coherent field emitter. From each point $\boldsymbol{r}_j$ within the emitting orbital, a spherical electron wave

$$\Psi_j(\boldsymbol{r}) \propto \frac{1}{|\boldsymbol{r}-\boldsymbol{r}_j|}\exp\left[i\left(\frac{2\pi\ |\boldsymbol{r}-\boldsymbol{r}_j|}{\lambda} + \varphi_j\right)\right] \quad (1)$$

emanates (Fig. 3d). Here $\lambda = h/[2m(qV - \Phi)]^{1/2}$ is the de Broglie wavelength of the electron, with mass m and elementary charge q, $qV - \Phi$ is

the kinetic energy of the electron just before it reaches the tip (Fig. 3a), h is Planck's constant and φ_j is the phase of the emitting orbital at point $\boldsymbol{r}_j$. Because all of these waves originate from the same orbital, they are coherent and interfere at the tip position $\boldsymbol{r}$, giving rise to an interference pattern as the tip is scanned in the detection plane. Simulations of the interference with phases φ_j of the LUMO of PTCDA (Fig. 3d) also produce a cloverleaf pattern (Fig. 3e), as in experiment. The additional concentric rings in Fig. 3c originate from field-emission resonances, as illustrated in Fig. 3f. Having passed the trapezoidal barrier, electrons may become transiently trapped between the metal tip and the barrier, because reflections at the tip and the barrier lead to the formation of resonance states. Because barrier tunnelling preserves coherence, these resonances lead to an oscillatory barrier transmission (Gundlach oscillations)[24]:

$$T_j(V,\boldsymbol{r}) \propto \frac{\exp(-B|\boldsymbol{r}-\boldsymbol{r}_j|)}{C+\sin(D|\boldsymbol{r}-\boldsymbol{r}_j|)} \quad (2)$$

where B, C and D are constants that depend on the applied voltage V and the work function Φ (see Methods).

Combining the effects of coherent electron emission from an extended orbital and Gundlach oscillations at a trapezoidal barrier, we modulate the amplitudes of the spherical waves $\Psi_j(\boldsymbol{r})$ with the barrier transmission $T_j(V, \boldsymbol{r})$, sum over all points $\boldsymbol{r}_j$ within the orbital from which electrons are emitted and calculate the square modulus $|\sum_j T_j(V,\boldsymbol{r})\Psi_j(\boldsymbol{r})|^2$. This yields the field-emission pattern shown in Fig. 3g, which is in excellent agreement with our experiment (Fig. 3c). This agreement confirms the successful fabrication of standing molecules and shows that the erected molecular structure keeps its vertical orientation even at extreme current densities of 10^8 A m^{-2}.

Cascading the charging of the quantum dot and field emission has several benefits. First, it enforces pre-emission coherence of the initial state across many atoms, thus enabling the realization of coherent emitters that consist of many more than one atom[25]. Owing to chemical perfection, in principle it is also possible to make emitters of arbitrary shapes, including fully coherent double-slit emitters[26]. Second, cascading permits the emission of single electrons on-demand[27], if combined with time-resolved gating[28]. Arbitrarily shaped on-demand emitters, arranged regularly on a surface, could be particularly promising for experiments with coherent electrons, including low-energy electron holography[29].

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0223-y.

Received: 5 February 2018; Accepted: 9 May 2018;
Published online 27 June 2018.

1. Eigler, D. M. & Schweizer, E. K. Positioning single atoms with a scanning tunnelling microscope. *Nature* **344**, 524–526 (1990).
2. Ternes, M., Lutz, C. P., Hirjibehedin, C. F., Giessibl, F. J. & Heinrich, A. J. The force needed to move an atom on a surface. *Science* **319**, 1066–1069 (2008).
3. Wagner, C. et al. Scanning quantum dot microscopy. *Phys. Rev. Lett.* **115**, 026101 (2015).
4. Stipe, B. C., Rezaei, M. A. & Ho, W. Single-molecule vibrational spectroscopy and microscopy. *Science* **280**, 1732–1735 (1998).
5. Heinrich, A. J., Gupta, J. A., Lutz, C. P. & Eigler, D. M. Single-atom spin-flip spectroscopy. *Science* **306**, 466–469 (2004).
6. Müllegger, S. et al. Radio frequency scanning tunneling spectroscopy for single-molecule spin resonance. *Phys. Rev. Lett.* **113**, 133001 (2014).
7. Baumann, S. et al. Electron paramagnetic resonance of individual atoms on a surface. *Science* **350**, 417–420 (2015).
8. Esat, T. et al. A chemically driven quantum phase transition in a two-molecule Kondo system. *Nat. Phys.* **12**, 867–873 (2016).
9. Loth, S., Etzkorn, M., Lutz, C. P., Eigler, D. M. & Heinrich, A. J. Measurement of fast electron spin relaxation times with atomic resolution. *Science* **329**, 1628–1630 (2010).
10. Saunus, C., Bindel, J. R., Pratzer, M. & Morgenstern, M. Versatile scanning tunneling microscopy with 120 ps time resolution. *Appl. Phys. Lett.* **102**, 051601 (2013).
11. Heinrich, A. J., Lutz, C. P., Gupta, J. A. & Eigler, D. M. Molecule cascades. *Science* **298**, 1381–1387 (2002).
12. Fuechsle, M. et al. A single-atom transistor. *Nat. Nanotechnol.* **7**, 242–246 (2012).
13. Donati, F. et al. Magnetic remanence in single atoms. *Science* **352**, 318–321 (2016).
14. Kalff, F. E. et al. A kilobyte rewritable atomic memory. *Nat. Nanotechnol.* **11**, 926–929 (2016).
15. Witte, G. & Wöll, C. Growth of aromatic molecules on solid substrates for applications in organic electronics. *J. Mater. Res.* **19**, 1889–1916 (2004).
16. Maurer, R. J. et al. Adsorption structures and energetics of molecules on metal surfaces: bridging experiment and theory. *Prog. Surf. Sci.* **91**, 72–100 (2016).
17. Eremtchenko, M., Schaefer, J. A. & Tautz, F. S. Understanding and tuning the epitaxy of large aromatic adsorbates by molecular design. *Nature* **425**, 602–605 (2003).
18. Temirov, R., Soubatch, S., Lucan, A. & Tautz, F. S. Free-electron-like dispersion in an organic monolayer film on a metal substrate. *Nature* **444**, 350–353 (2006).
19. Schuler, B. et al. Adsorption geometry determination of single molecules by atomic force microscopy. *Phys. Rev. Lett.* **111**, 106103 (2013).
20. Toher, C. et al. Electrical transport through a mechanically gated molecular wire. *Phys. Rev. B* **83**, 155402 (2011).
21. Green, M. F. et al. Patterning a hydrogen-bonded molecular monolayer with a hand-controlled scanning probe microscope. *Beilstein J. Nanotechnol.* **5**, 1926–1932 (2014).
22. Kocić, N. et al. Periodic charging of individual molecules coupled to the motion of an atomic force microscopy tip. *Nano Lett.* **15**, 4406–4411 (2015).
23. Forbes, R. G. Physics of generalized Fowler–Nordheim-type equations. *J. Vac. Sci. Technol. B* **26**, 788–793 (2008).
24. Gundlach, K. Zur Berechnung des Tunnelstroms durch eine trapezförmige Potentialstufe. *Solid-State Electron.* **9**, 949–957 (1966).
25. Fink, H. W. Mono-atomic tips for scanning tunneling microscopy. *IBM J. Res. Develop.* **30**, 460–465 (1986).
26. Oshima, C. et al. Young's interference of electrons in field emission patterns. *Phys. Rev. Lett.* **88**, 038301 (2002).
27. Fève, G. et al. An on-demand coherent single electron source. *Science* **316**, 1169–1172 (2007).
28. Cocker, T. L., Peller, D., Yu, P., Repp, J. & Huber, R. Tracking the ultrafast motion of a single molecule by femtosecond orbital imaging. *Nature* **539**, 263–267 (2016).
29. Longchamp, J. N. et al. Imaging proteins at the single-molecule level. *Proc. Natl Acad. Sci. USA* **114**, 1474–1479 (2017).
30. Weiß, S. et al. Exploring three-dimensional orbital imaging with energy-dependent photoemission tomography. *Nat. Commun.* **6**, 8287 (2015).

Acknowledgements We thank L. Kronik and S. Sarkar (Weizmann Institute of Science) and M. Rohlfing (Universität Münster) for performing DFT calculations of standing molecules (not reported here).

Reviewer information *Nature* thanks T. Greber, A. Heinrich and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.E., F.S.T. and R.T. conceived the research. T.E., N.F. and R.T. conducted the NC-AFM/STM experiments and analysed the resultant experimental data. T.E., R.T. and F.S.T. interpreted the data. T.E. carried out simulations and prepared the figures. T.E., R.T. and F.S.T. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0223-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to F.S.T.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample preparation. An atomically clean Ag(111) surface was prepared in ultra-high vacuum by repeated cycles of Ar^+ sputtering and annealing at 800 K. A small coverage of PTCDA molecules (less than 20% of a monolayer) was deposited onto the clean Ag(111) surface, held at room temperature, from a custom-built Knudsen cell. After deposition, the sample was moved into the microscope and cooled to 5 K. Heating a thin silver wire, we evaporated silver atoms onto the Ag(111) surface in situ at $T \approx 10$ K, thus preventing clustering.

Tip preparation. We used a PtIr wire as the SPM tip, sharpened with a focused-ion beam (FIB). The PtIr tip was treated in situ by applying controlled voltage pulses and indentations into the clean Ag surface, resulting in a clean Ag-covered tip. The cleanness of the tip was validated by the spectroscopic signature of the Ag(111) surface state and STM imaging of the LUMO of PTCDA. For the high-resolution AFM imaging shown in Fig. 1c, the tip was functionalized with a carbon monoxide molecule[31].

NC-AFM/STM experiments. The experiments were carried out in a non-contact atomic force/scanning tunnelling microscope (NC-AFM/STM; CREATEC) under ultrahigh-vacuum conditions at a temperature of $T \approx 5$ K. The microscope is equipped with a qPlus sensor (CREATEC) and operated in frequency-modulation mode (resonance frequency $f_0 \approx 31$ kHz, quality factor $Q \approx 4 \times 10^4$, spring constant $k = 1{,}800$ N m^{-1}, oscillation amplitude $A \approx 0.2$–0.3 Å). All bias voltages quoted here were applied to the sample while the tip was grounded. STM images were acquired in constant-current mode. The AFM images were recorded in constant-height mode at $V = 2$ mV and show the frequency shift $\Delta f(z) \approx -[f_0/(2k)]\mathrm{d}F_z/\mathrm{d}z$. Field-emission images were recorded in constant-height mode with qPlus oscillations turned off, and show the current I.

Structure of the PTCDA + 2Ag complex. Density functional theory (DFT) calculations were used to determine the structure of the PTCDA + 2Ag complex in the gas phase. They were carried out with GAMESS[32] using a 3–21 G basis set including polarization functions for heavy atoms and the hybrid Becke three-parameter Lee–Yang–Parr (B3LYP) functional. The structure of the standing molecules shown in Fig. 1 and Extended Data Fig. 1 is the gas-phase structure of the PTCDA + 2Ag complex.

Tilting stiffness of the standing molecule. When moving the tip towards the standing molecule along the y direction (Fig. 2a, b), the molecule cannot succumb to the attractive force by tilting towards the tip, because this would require breaking one of the Ag–O bonds at the surface. The F_y force map shown in the bottom panel of Fig. 2b therefore reveals the intrinsic range of the attractive force between the tip and the molecule. By contrast, when the tip is moved along the x direction (top panel of Fig. 2b), the molecule is able to tilt towards the tip. This leads to the more prolate features in the F_x map. Comparing the F_x and F_y maps in Fig. 2b, we can estimate the force constant for the initial tilting of the molecule out of the upright orientation.

In detail, this proceeds as follows. We model the force between the tip and a corner oxygen atom, which is the most reactive part of the molecule, as

$$\boldsymbol{F}(r) = A\mathrm{e}^{-r/\rho}\boldsymbol{e}_r \tag{3}$$

where r is the distance between a corner oxygen atom and the tip apex, $\boldsymbol{e}_r$ the unit vector in this direction and ρ an appropriate decay constant. For the approach along the y direction, we know that the molecule cannot bend towards the tip. When calculating the force acting from the tip onto the molecule, we take into account that there are two corner oxygen atoms, separated by $2a$ along the y axis, for each of which the force law in equation (3) holds. We therefore obtain for the lateral component of the force onto the molecule

$$F_y(y) = A[\mathrm{e}^{-d/(\rho\sin\alpha_1)}\cos\alpha_1 + \mathrm{e}^{-d/(\rho\sin\alpha_2)}\cos\alpha_2] \tag{4}$$

where d is the vertical distance between the top of the molecule and the imaging plane, $\alpha_1 = \arctan[d/(y+a)]$ and $\alpha_2 = \arctan[d/(y-a)]$ (Extended Data Fig. 4a). Equation (4) can be used to fit the experimental $F_y(y)$ profile in Extended Data Fig. 4c (black data points), extracted from the bottom panel of Fig. 2b, using the amplitude A and the decay length ρ as fit parameters. As the blue curve in Extended Data Fig. 4c shows, we obtain an excellent fit for $\rho = 119$ pm.

In the next step, we use the force law of equation (3) to extract the bending stiffness of the standing molecule when the tip approaches along x. The $F_x(x)$ profile, extracted from the top panel of Fig. 2b, is shown in Extended Data Fig. 4c as red data points. This force corresponds to an equilibrium between the lateral attraction $F_x(x)$, which acts between the tip and the molecule, tilting the molecule towards the tip, and the lateral component $F_{\theta,x}(x)$ of the restoring force that tries to bring the molecule back into the vertical (see Extended Data Fig. 4b). For each x, we therefore have to find the tilting angle $\theta_{\mathrm{eq}}(x)$ for which the force equilibrium

$$|F_x(x, \theta_{\mathrm{eq}}(x))| = |F_{\theta,x}(x, \theta_{\mathrm{eq}}(x))| \tag{5}$$

holds. For the restoring force F_θ, we assume Hook's law for the corresponding torque, $lF_\theta \approx \kappa_\theta\theta$, where l is the length of the standing molecule and κ_θ is the torsional spring constant. Its x component therefore becomes $F_{\theta,x} \approx (\kappa_\theta\theta/l)\cos\theta$. The total force on the molecule, $F_{\mathrm{total}} = |\boldsymbol{F}_1 + \boldsymbol{F}_2|$ in Extended Data Fig. 4b, is then

$$F_{\mathrm{total}} = 2|\boldsymbol{F}(r)|\sin\gamma \tag{6}$$

and by symmetry located in the x–z plane. From Extended Data Fig. 4b, we obtain the following geometric relations:

$$\begin{aligned} \cos\gamma &= \frac{a}{r} \\ r &= \sqrt{(x - x_{\mathrm{m}})^2 + a^2 + b^2} \\ x_{\mathrm{m}} &= l\sin\theta \\ b &= s + d = (x - x_{\mathrm{m}})\tan\alpha \end{aligned} \tag{7}$$

where γ, α, s, b and x_{m} are defined in Extended Data Fig. 4b. Projecting F_{total} (equation (6)) onto the x direction, we obtain

$$F_x = 2|\boldsymbol{F}(r)|\sin\gamma\cos\alpha \tag{8}$$

With the help of equation (7), F_x can be expressed as a function of (x, θ). For each tip position x, the tilting angle θ is then adjusted to fulfil the equilibrium condition (equation (5)). The results are shown in Extended Data Fig. 4c–e, in which the left panels display the simulated lateral force F_x (green curve), in comparison with the experimental F_x (red data points) and F_y (black data points); the corresponding tilt angles $\theta_{\mathrm{eq}}(x)$ and linear elongations $x_{\mathrm{m}}(x)$ are plotted in the middle and right panels. Extended Data Fig. 4c shows the best fit, yielding an initial tilting stiffness of $\kappa_\theta = 630$ zN m rad^{-1}, corresponding to $\kappa = \kappa_\theta/l^2 = 0.38$ N m^{-1}. For comparison, the stiffness of a carbon monoxide molecule at the tip of a SPM is $\kappa = 0.24$ N m^{-1}, whereas on a Cu(111) surface it is $\kappa = 1.64$ N m^{-1} (ref. [33]). Extended Data Fig. 4d shows that for large tilting stiffnesses the simulated F_x for the approach along x converges to the measured F_y for approach along y, as it must. Finally, Extended Data Fig. 4e illustrates the effect of a soft restoring-force constant on F_x and the molecular tilt θ_{eq} or elongation x_{m}. We note that, owing to anharmonicity in F_θ and other secondary effects, our model is applicable for small x only.

Structure at the interface between the standing molecule and the Ag(111) surface. By combining translational and rotational manipulation of the standing molecule, we can determine the structure at the interface between the standing molecule and the Ag(111) surface.

A single translational manipulation step is illustrated in Extended Data Fig. 3a. By repeating this procedure several times, the standing molecule is moved along a specific trajectory on the surface. Part of this trajectory is plotted in the AFM image in Fig. 2c. Analysing the trajectory, we find that possible adsorption sites of the standing molecule form the same lattice as the hollow sites on the Ag(111) surface (Extended Data Fig. 6). This suggests that the silver atoms of the standing molecule sit in hollow sites. Rotational manipulation data verify this conjecture.

A single rotational manipulation step is illustrated in Extended Data Fig. 3b. The standing molecule tends to rotate towards the tip position. Analysing a large number of erected and/or rotated molecules (Extended Data Fig. 5), we find that standing molecules exist only in six distinct lateral orientations on the Ag(111) surface. These are indicated by the three black lines and three red lines in Fig. 2d. We know that the distance between the silver atoms in the standing molecule is approximately 2.9 Å, because in the AFM images of the standing molecule this is the distance between the maxima of the attractive force, which acts between the oxygen atoms at the upper end of the standing molecule and the silver atom at the apex of the tip (Fig. 1e). Because this Ag–Ag distance fits well to distances between the hollow sites of the Ag(111) surface, which amount to 2.9 Å (red squares in Fig. 2d) and 3.4 Å (black circles), we are led to the conclusion that the silver atoms of the standing molecule do indeed sit in hollow sites. Evidently, the epitaxial registry between the essentially rigid Ag–O arrangement in the standing molecule and the hollow sites of the Ag(111) surface determines the possible orientations of the standing molecule. Note that in AFM images molecules in the orientation shown in black in Fig. 2d appear asymmetric (for example, in Extended Data Fig. 3b after manipulation), because the mismatch between the silver atoms in the standing molecule and the hollow sites in this surface direction enforces a tilt of the standing molecule.

Altogether, we have built and erected at least 107 different PTCDA molecules. We measured the orientation of only some of them. To obtain a larger dataset regarding the orientation of standing molecules, we rotated several standing molecules, as demonstrated in Extended Data Fig. 3b. The total number of measured orientations is therefore 128 (Extended Data Fig. 5). However, to avoid any directionality during the collection of rotation statistics, we placed the tip above the centre of the standing molecule (not next to it, as in Extended Data Fig. 3b). By this procedure the standing molecule rotates randomly.

The two Ag–Ag distances of 2.9 Å and 3.4 Å also correspond to different types of hollow site of the Ag(111) surface. A spacing of 2.9 Å coincides with the distance between two adjacent hexagonal close packed (hcp) or two adjacent face-centred

cubic (fcc) sites (red squares in Fig. 2d), whereas for a spacing of 3.4 Å one of the hollows is an hcp site and the other is an fcc site. We find that the fcc/hcp configuration (black circles) is less abundant than the fcc/fcc or hcp/hcp configuration (red squares): 52 versus 76 out of a total of 128 (see Extended Data Fig. 5). This is explained by the better fit of the fcc/fcc and hcp/hcp configurations to the natural distance of two silver atoms bound to the dianhydride group of PTCDA (2.9 Å) and the higher degeneracy of these configurations.

A distinction between fcc/fcc and hcp/hcp configurations is not possible in either AFM or the field-emission experiment. The field-emission images in Fig. 3c correspond to different adsorption configurations, because a rotational step of 30° from the 30° to the 60° orientation (Fig. 2d), as seen in the field-emission images, must correspond to a change from an fcc/hcp to an hcp/hcp or fcc/fcc configuration (Fig. 2d). However, the interference patterns look identical within experimental accuracy, showing that the effect of different adsorption configurations on the emitted electrons is negligible.

Stabilization mechanism of the upright configuration. Manipulation experiments on the standing molecule reveal elements of the stabilization mechanism of the upright configuration. We observe that a positive bias voltage applied to the surface topples the molecule over (Extended Data Fig. 3c), suggesting that the standing molecule is negatively charged. Similarly, if two standing molecules are brought very close to each other, they both topple over. This observation is explained either by the electrostatic force between the standing molecules or by the mutual destabilization of the charge on each of them. The fact that the standing molecule acts as a quantum dot with integer charge states (Fig. 3a) suggests that the stability of its upright configuration may result from electron–electron correlations. This would explain why DFT calculations do not find a metastable upright configuration for the PTCDA + 2Ag complex on Ag(111) and sets our case apart from others, where upright standing configurations are stabilized by intermolecular interactions[15,34–36] or directional bonds to the surface[37–40].

Detection of the field-emission current. We detect the field-emission current with the tip of our STM/AFM. Unlike in tunnelling experiments, the current is detected over the complete surface of the tip. Specifically, the field-emission current does not distinguish between electrons that hit the tip apex (well-defined path length) and all other positions on the tip surface (no well-defined path length). The sum of the latter contributes to the large background in the field-emission current. The background that arises from our mode of detecting the field emission is thus not necessarily incoherent, although other mechanisms that create decoherence may be present. The signal-to-background ratio is tip-dependent, and approximately 2% in the data in Fig. 3b, c and Extended Data Fig. 7.

Modelling the interference pattern. In accordance with ref. [24], we used the following expressions in equation (2) for the barrier transmission $T_j(V, \boldsymbol{r})$ for electrons at the Fermi level:

$$B = \frac{4}{3}\frac{\sqrt{2m}}{\hbar}\frac{\Phi^{3/2}}{qV}$$

$$C = 2\frac{qV-\Phi}{\Phi+E_F}+1$$

$$D = \frac{4}{3}\frac{\sqrt{2m}}{\hbar}\frac{(qV-\Phi)^{3/2}}{qV}$$

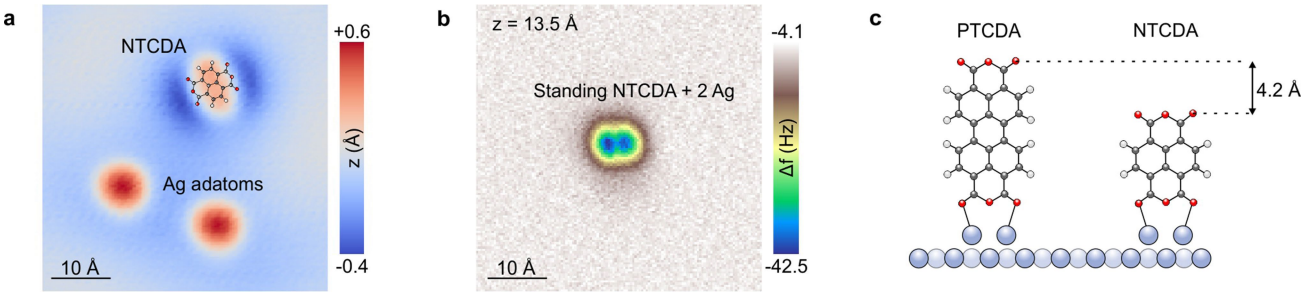
The work function Φ of Ag(111) was taken to be equal to 4.4 eV[41]. When simulating the field-emission images in Fig. 3e–h, grid points $\boldsymbol{r}_j$ within an isosurface that contains 90% of the charge density of the orbital were used.

In equation (1), we have approximated the de Broglie wavelength λ of the field-emitted electron by its value just before it reaches the tip. In reality, its energy and therefore also λ change on its passage towards the tip. This issue has been considered previously[26], and it was shown that the effect of the electron acceleration should not induce substantial changes in the interference patterns observed. We validated this directly by collecting field-emission images at different voltages (Extended Data Fig. 7). Gundlach transmission in equation (2) takes the acceleration of the electrons into account, because it is calculated explicitly for a trapezoidal barrier. Hence, the fine structure of the field-emission pattern that originates from Gundlach resonances is not affected by the above approximation.

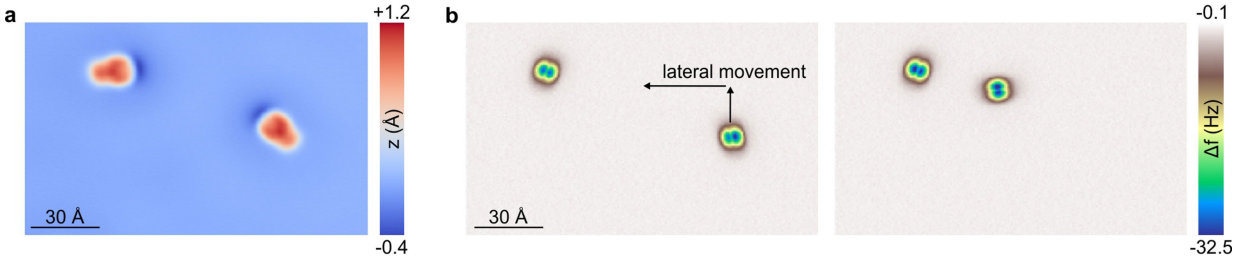
The barrier transmission in equation (2) has an exponential dependence on the distance between the emission and detection points. Thus, the contributions from the lower parts of the emitting orbital to the interference pattern (Fig. 3e–h) are negligible. Nevertheless, the amplitude and phase distributions that we observe for the upper part of the molecule can exist only if the state of the electron before emission is coherent over all 38 atoms of the molecule, because the molecular orbital from which the electrons are emitted is extended over the complete molecule, regardless whether emission from the lower part is negligible.

Data availability. The datasets generated and analysed during this study are available from the corresponding author on reasonable request.

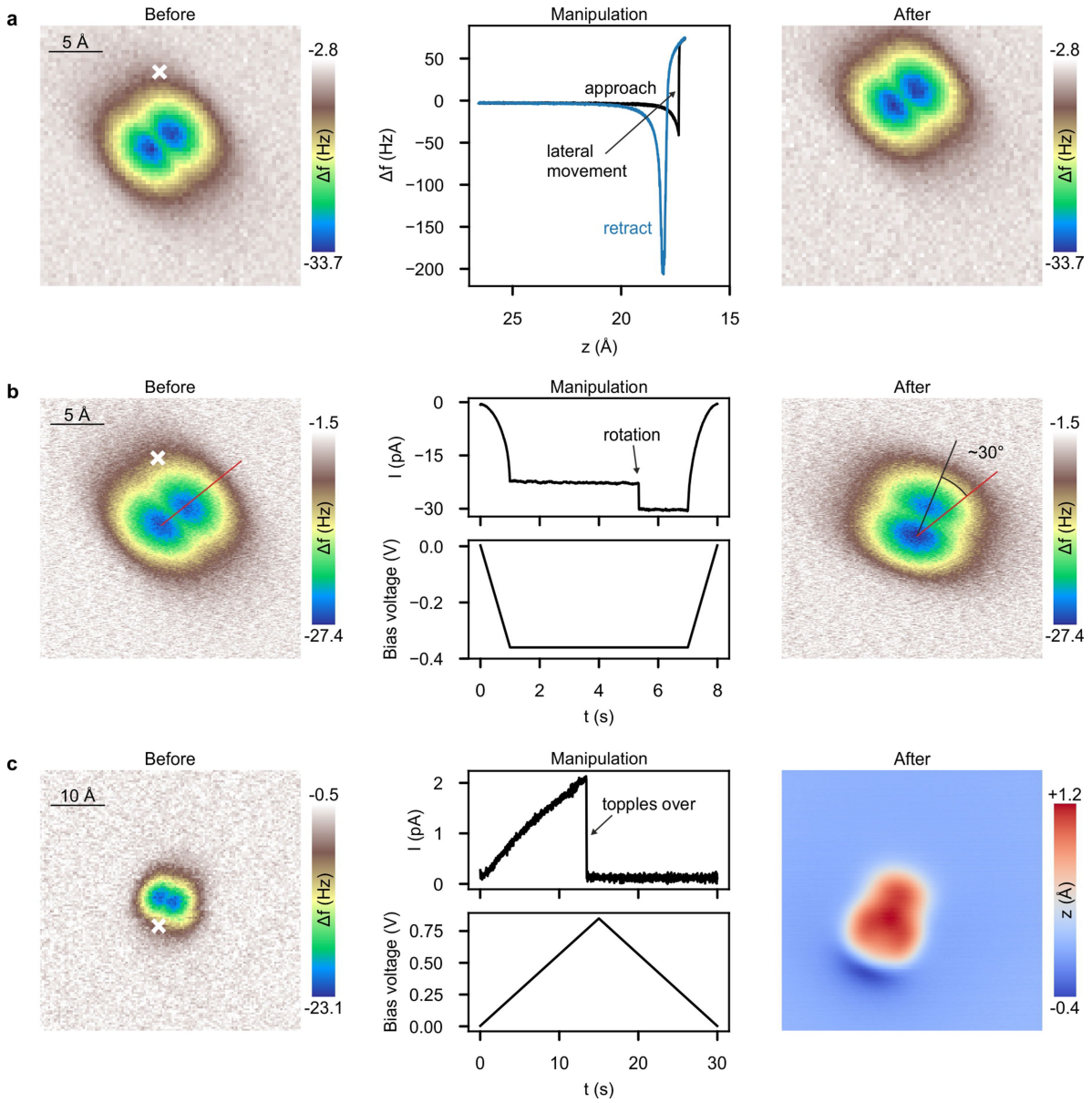
31. Gross, L., Mohn, F., Moll, N., Liljeroth, P. & Meyer, G. The chemical structure of a molecule resolved by atomic force microscopy. *Science* **325**, 1110–1114 (2009).
32. Schmidt, M. W. et al. General atomic and molecular electronic structure system. *J. Comput. Chem.* **14**, 1347–1363 (1993).
33. Weymouth, A. J., Hofmann, T. & Giessibl, F. J. Quantifying molecular stiffness and interaction with lateral force microscopy. *Science* **343**, 1120–1122 (2014).
34. Demuth, J. E., Christmann, K. & Sanda, P. N. The vibrations and structure of pyridine chemisorbed on Ag(111): the occurrence of a compressional phase transformation. *Chem. Phys. Lett.* **76**, 201–206 (1980).
35. Lee, I., Son, S., Shin, T. & Hahn, J. R. Direct observation of the conformational transitions of single pyridine molecules on a Ag(110) surface induced by long-range repulsive intermolecular interactions. *J. Chem. Phys.* **146**, 014706 (2017).
36. Ulman, A. Formation and structure of self-assembled monolayers. *Chem. Rev.* **96**, 1533–1554 (1996).
37. Blyholder, G. Molecular orbital view of chemisorbed carbon monoxide. *J. Phys. Chem.* **68**, 2772–2777 (1964).
38. Cai, Y., Guo, Y., Xu, X. & Jiang, B. First-principle investigation 3,4-ethylenedioxythiophene molecule adsorption on Cu(110)-(2 × 1)O surface. *Surf. Sci.* **665**, 83–88 (2017).
39. Jasper-Tönnies, T. et al. Conductance of a freestanding conjugated molecular wire. *Phys. Rev. Lett.* **119**, 066801 (2017).
40. Gerhard, L. et al. An electrically actuated molecular toggle switch. *Nat. Commun.* **8**, 14672 (2017).
41. Chelvayohan, M. & Mee, C. H. B. Work function measurements on (110), (100) and (111) surfaces of silver. *J. Phys. C* **15**, 2305–2312 (1982).



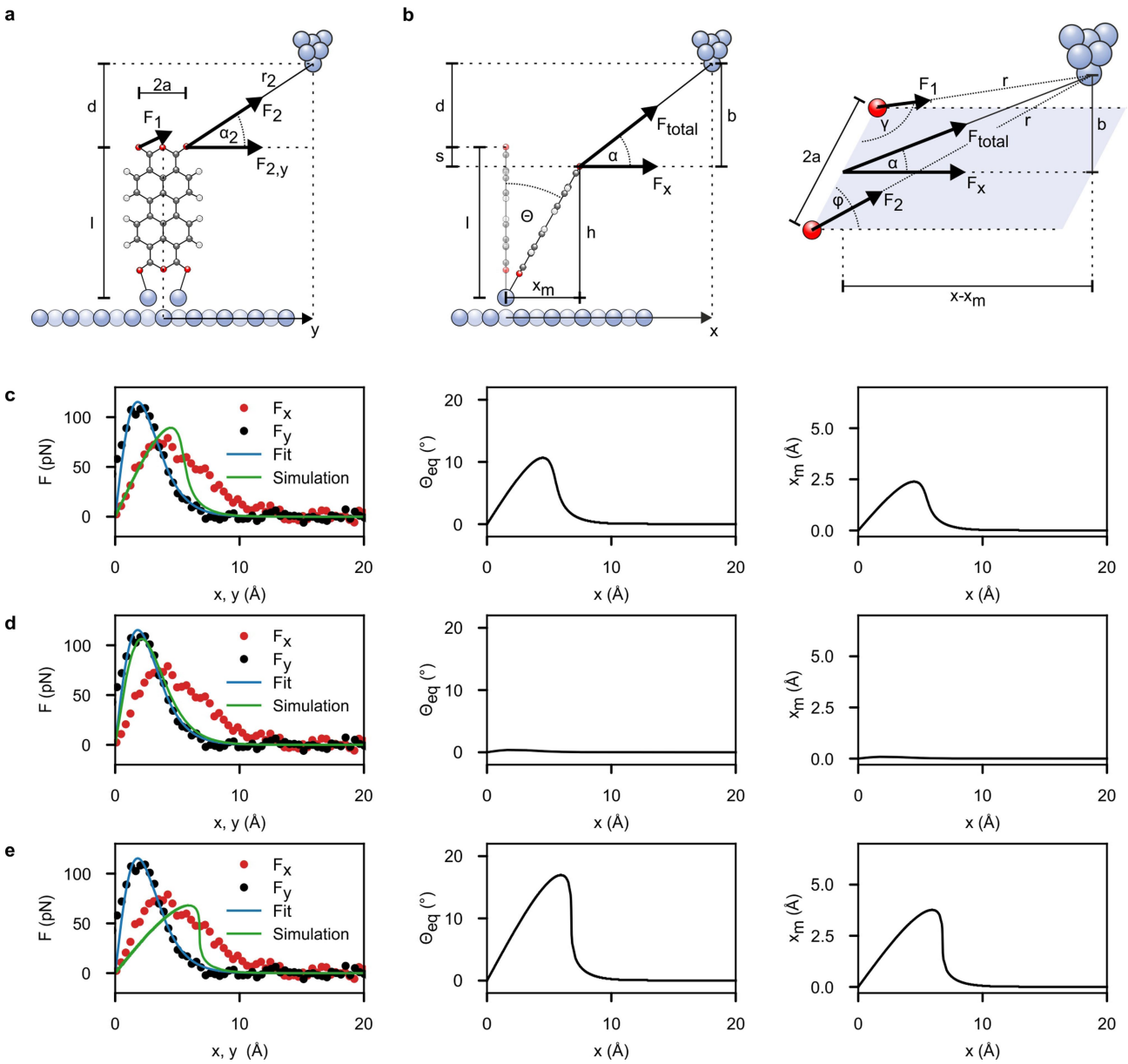
Extended Data Fig. 1 | A standing NTCDA molecule. a, Constant-current STM image of an NTCDA (1,4,5,8-naphthalenetetracarboxylic-dianhydride) molecule and two silver adatoms, recorded before the assembly of a NTCDA + 2Ag complex. **b**, AFM image of a standing NTCDA molecule, recorded at a tip height of $z = 13.5$ Å above the surface. **c**, Schematic side view of standing PTCDA and NTCDA molecules. The length difference of 4.2 Å between the two molecules corresponds well with the tip-height difference $\Delta z = 17.5$ Å − 13.5 Å = 4.0 Å between the AFM images of the standing PTCDA (see Fig. 1f) and NTCDA (**b**) molecules.



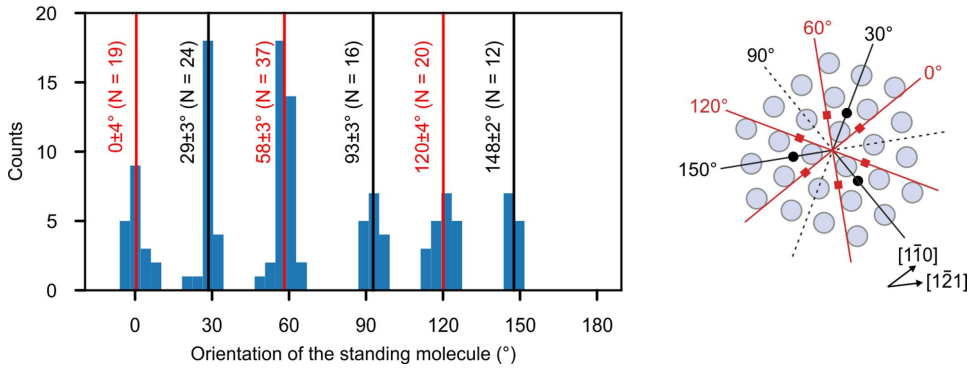
Extended Data Fig. 2 | Two standing PTCDA molecules. a, Constant-current STM image of two PTCDA + 2Ag complexes that were assembled in the same way as shown in Fig. 1b. **b**, AFM image of the standing molecules, recorded at tip height of $z = 17.5$ Å above the surface (left). One of the standing molecules is then moved closer to the other by the lateral manipulation procedure demonstrated in Extended Data Fig. 3a (right).



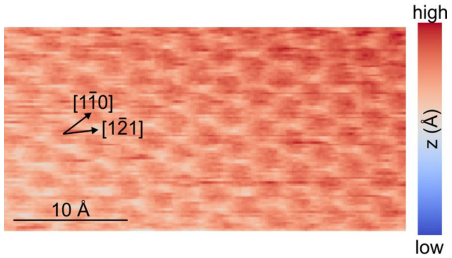
Extended Data Fig. 3 | Manipulation of the standing molecule. a, Lateral movement of the standing molecule by tip approach. The white cross in the AFM image on the left indicates the tip position during manipulation ($V = 2$ mV). **b**, Rotational movement of the standing molecule by a current pulse. The white cross in the AFM image on the left indicates the tip position during manipulation ($z = 17.5$ Å). The molecule jumps from a red (symmetric) to a black (asymmetric) position (see Methods). **c**, Toppling over the standing molecule to the surface by using a positive bias-voltage sweep. The white cross in the AFM image on the left indicates the tip position during manipulation ($z = 17.5$ Å). The constant-current STM image on the right shows the PTCDA + 2Ag complex after the toppling.



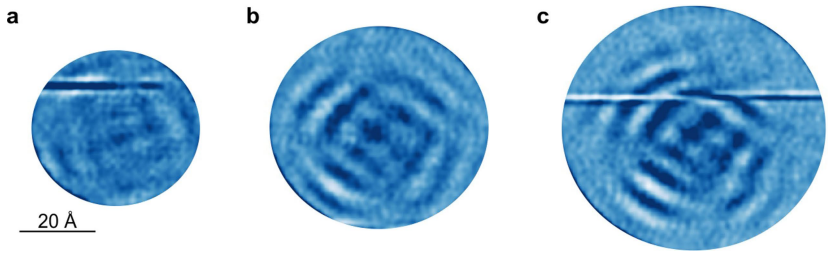
Extended Data Fig. 4 | Determining the tilting stiffness of the standing molecule. a, Approach along the y direction as defined in Fig. 2a, b. Parameters are $d = 3.25$ Å, $l = 12.9$ Å and $2a = 4.55$ Å; $d + l \neq z$, because the tip height z is measured from the centre of the uppermost surface layer, whereas l is measured from the centre of the two silver adatoms in the PTCDA + 2Ag complex. **b**, Approach along the x direction as defined in Fig. 2a, b. The sketch on the left shows a side view, whereas on the right a perspective view onto the top of the molecule is drawn. **c**, Left, best fit (green line) of the experimental F_x (on x axis, red circles), obtained with the model in equation (8) and $\kappa_\theta = 630$ zN m rad^{-1} ($\kappa = 0.38$ N m^{-1}). Black data points display F_y (on y axis), fitted with equation (4) (blue line). **d**, As in **c**, but for a simulated curve (green) with $\kappa_\theta = 20.0$ aN m rad^{-1} ($\kappa = 12.02$ N m^{-1}), which is too stiff to reproduce the experimental F_x (red). **e**, As in **c**, but for a simulated curve (green) with $\kappa_\theta = 310$ zN m rad^{-1} ($\kappa = 0.19$ N m^{-1}), which is too soft. For a detailed discussion of this figure, see Methods. Tilt angles θ_{eq} and linear elongations x_m are plotted in the middle and right panels in **c**–**e**.



Extended Data Fig. 5 | Orientation of the standing molecule. The histogram (left) shows all evaluated orientations of standing PTCDA molecules on the Ag(111) surface. The angle is defined as in Fig. 2d and Extended Data Fig. 3b. In total, we evaluated the orientations of 128 standing PTCDA molecules. The possible orientations on the Ag(111) surface are illustrated on the right. Black and red symbols indicate the possible positions of one of the silver atoms at the surface contact, when the other sits in the centre. See also Fig. 2d and Methods.



Extended Data Fig. 6 | The Ag(111) lattice. Constant-current STM image of the atomically resolved Ag(111) surface. In all experiments, the Ag(111) lattice orientation was as shown in this image.



Extended Data Fig. 7 | Field-emission images. a–c, Successive field-emission images (without the background) recorded at $z = 73.5$ Å and bias voltages of $V = -24.00$ V (**a**), $V = -24.35$ V (**b**) and $V = -24.70$ V (**c**).

LETTER

https://doi.org/10.1038/s41586-018-0221-0

Self–assembly of highly symmetrical, ultrasmall inorganic cages directed by surfactant micelles

Kai Ma[1], Yunye Gong[2], Tangi Aubert[1,3], Melik Z. Turker[1], Teresa Kao[1], Peter C. Doerschuk[2,4] & Ulrich Wiesner[1]*

Nanometre-sized objects with highly symmetrical, cage-like polyhedral shapes, often with icosahedral symmetry, have recently been assembled from DNA[1–3], RNA[4] or proteins[5,6] for applications in biology and medicine. These achievements relied on advances in the development of programmable self-assembling biological materials[7–10], and on rapidly developing techniques for generating three-dimensional (3D) reconstructions from cryo-electron microscopy images of single particles, which provide high-resolution structural characterization of biological complexes[11–13]. Such single-particle 3D reconstruction approaches have not yet been successfully applied to the identification of synthetic inorganic nanomaterials with highly symmetrical cage-like shapes. Here, however, using a combination of cryo-electron microscopy and single-particle 3D reconstruction, we suggest the existence of isolated ultrasmall (less than 10 nm) silica cages ('silicages') with dodecahedral structure. We propose that such highly symmetrical, self-assembled cages form through the arrangement of primary silica clusters in aqueous solutions on the surface of oppositely charged surfactant micelles. This discovery paves the way for nanoscale cages made from silica and other inorganic materials to be used as building blocks for a wide range of advanced functional-materials applications.

In our search for dodecahedral silica-cage structures (Fig. 1), we began with the early stages in the self-assembly of silica structures directed by surfactant micelles[14]. Our synthesis system contained cetyltrimethylammonium bromide (CTAB) surfactant micelles and tetramethyl orthosilicate (TMOS) as a sol–gel silica precursor (see Methods). We added hydrophobic mesitylene (TMB) into the aqueous CTAB micelle solution, increasing the size and deformability of the micelles[15]. We selected TMOS as the silica source because of its fast hydrolysis rate in water, and we adjusted the initial reaction pH to about 8.5. When TMOS was added, its hydrolysis to silicic acid reduced the reaction pH to neutral[16,17]. The lowered pH accelerated silane condensation, leading to primary silica clusters[18] with diameters of around 2 nm. The negatively charged silica clusters were attracted to the positively charged CTAB micelle surface, assembling into micelle-templated nanostructures[14]. This experimental design—in which fast hydrolysis and condensation of the silica precursor quickly terminated the reaction—allowed the early stages in the micelle-directed self-assembly of silica structures to be preserved[17].

In order to improve the dispersity of particles on transmission electron microscopy (TEM) grids, we added low-molar-mass, silane-modified monofunctional polyethylene glycol (PEG) into the solution one day before preparing TEM samples. This process covalently coated the accessible silica surface[19], yielding PEGylated nanoparticles (Extended Data Fig. 1) that could be further purified and isolated from the synthesis solution. Using TEM, we observed particles of narrow size distribution, with an average diameter of around 12 nm (Fig. 2a, including inset), consistent with silica structures wrapped around TMB-swollen CTAB micelles[15]. The detailed particle structure was difficult to identify, however. Therefore, we plasma-etched TEM samples on carbon grids for five seconds before imaging in order to remove excess organic chemicals (such as PEG-silane) that would otherwise contribute to background noise. To further improve the signal-to-noise ratio, we acquired and averaged a series of images of the same sample area. Stripes and windows in zoomed-in images of individual particles became more clearly recognizable, suggesting the presence of cage-like structures (Fig. 2b, including insets).

Analysis of thousands of such single-particle TEM images revealed the prevalence of two cage projections, one with two-fold symmetry and one with five-fold symmetry (Fig. 2c)—too few to allow a successful 3D reconstruction. We therefore shifted our attention to cryo-electron microscopy (cryo-EM) characterization of the native reaction solution. We omitted the silica-surface PEGylation step, as the high PEG concentration substantially increased the sensitivity of the samples to radiation, resulting in difficulties in obtaining clear cryo-EM images.

Cryo-EM provided direct visualization of particles in solution with arbitrary orientation—that is, without disturbances resulting from drying of the samples on TEM substrates, including structure deflation. The background noise was much reduced because of the absence of a TEM substrate and of chemicals dried onto the substrate during sample preparation (Fig. 2c). Although particle aggregation was observed occasionally through cryo-EM (Fig. 2d), individual silica nanoparticles with cage-like structures could always be identified (Fig. 2c, d). No particle aggregation was observed in dry-state TEM of PEGylated particles, suggesting that particle aggregation observed by cryo-EM was a reversible process that could be overcome via insertion of PEG chains.

We manually identified around 19,000 single-particle images from cryo-EM micrographs, then clustered them and averaged the images in each cluster in order to improve the signal-to-noise ratio[20]. The averages showed different orientations of silica nanoparticles with cage-like structures (silicages; Extended Data Fig. 2a). We identified averages that were consistent with selected projections of a pentagonal dodecahedral cage (Fig. 2c and Extended Data Fig. 2b). The dodecahedral silicage (icosahedral point group, I_h; Fig. 1) is the simplest of a set of Voronoi polyhedra suggested to form the smallest structural

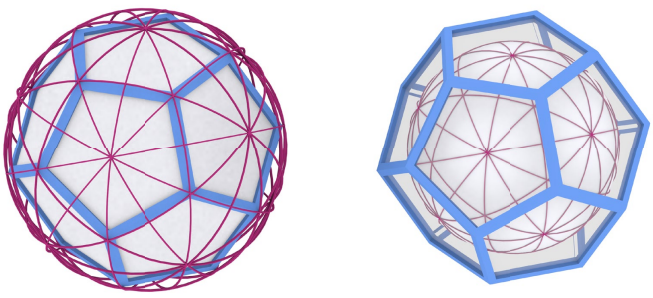


Fig. 1 | Representations of a dodecahedron. Left, among the platonic solids, the dodecahedron best fills out its circumscribed sphere—that is, a sphere that passes through all of its vertices. Right, the inscribed sphere that passes through all of the dodecahedron's facets is shown for comparison.

[1]Department of Materials Science and Engineering, Cornell University, Ithaca, NY, USA. [2]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA. [3]Department of Chemistry, Ghent University, Ghent, Belgium. [4]Nancy E. and Peter C. Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA. *e-mail: ubw1@cornell.edu

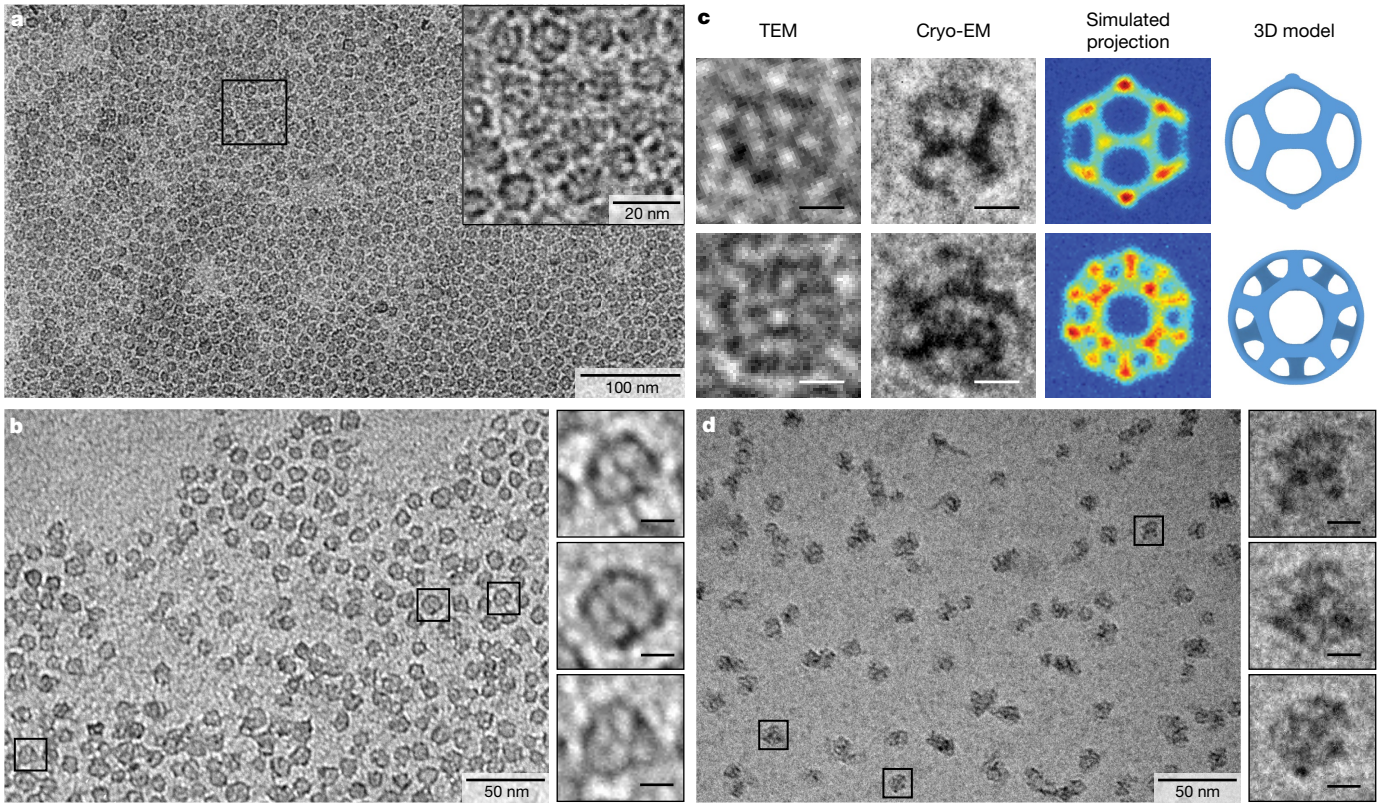


Fig. 2 | TEM and cryo-EM characterizations of silicages. a, Low-magnification TEM images of PEG-coated silicages on a carbon substrate. The inset is a zoomed-in image. **b**, Averaged TEM image obtained using 11 images of the same sample area of PEG-coated silicages; insets show representative individual structures at a higher magnification. The sample was plasma etched for five seconds before TEM characterization to reduce background noise. **c**, Comparison of silicages observed by TEM and cryo-EM with projections of simulated dodecahedral cages and models. **d**, Cryo-EM images of silicages without PEG coating. Scale bars in the insets of **b**–**d** represent 5 nm.

units of multiple forms of mesoporous silica[21]. Although such highly symmetrical ultrasmall silica cages have, to our knowledge, never been isolated before, it had seemed likely that this should be possible.

Guided by this structural insight, we carried out single-particle 3D reconstructions of silicages using the 'Hetero'[22] model-based maximum likelihood machine-learning algorithm, in which a two-class reconstruction is computed to overcome challenges associated with structural heterogeneity (see Methods) and rotational icosahedral symmetry is imposed on both classes (Extended Data Fig. 2). One of the two-class reconstructions was a dodecahedral cage, as visualized by UCSF Chimera (ref. [23]; Fig. 3a, b). We identified a low-intensity signal inside this reconstructed cage, consistent with the presence of TMB-swollen CTAB micelles within the silicage, whose electron density is lower than that of silica but higher than that of the surrounding ice. The other class of reconstruction did not provide an interpretable structure, probably owing to heterogeneity in the structure of the corresponding particles. We carried out such two-class reconstructions using different numbers of single-particle images (2,000, 7,000 and 10,000) and found consistent results. We further performed single-class reconstructions with the Hetero algorithm, using only images from the class that showed dodecahedral cages in two-class reconstructions. Equivalent two-class and single-class reconstructions were also performed with the widely used RELION 2.1 system[24,25]. Dodecahedral cage structures were obtained in all of these reconstructions (Extended Data Fig. 3 and Supplementary Videos 1, 2). The resolution of the reconstructions was approximately 2 nm (ref. [26]; Extended Data Fig. 4). Silica in these cages is amorphous at the atomic level, which prevented atomic resolution of these reconstructions.

The Hetero reconstruction algorithm provided estimates of the projected orientation (that is, three Euler angles) for each experimental image, which we used to compute predicted projections. Nine predicted projections and corresponding experimental images were manually clustered, and averages were computed for each cluster (Fig. 3c). The similarity of the projections of the 3D reconstruction and the averaged experimental images supports the dodecahedral cage structure. Furthermore, we calculated the theoretical probabilities of finding each of the nine projections (Fig. 3c) on the basis of the assumption that the orientations of silicages in cryo-EM are random. We then compared the results with the probabilities observed by single-particle 3D reconstruction (Extended Data Fig. 5). The high consistency between the theoretical and experimental projection probabilities further supports the dodecahedral cage reconstruction.

At this early point we can only speculate about the exact mechanism by which the observed silicage structure is formed, but there are clues in the details of the reconstruction. The silica occupying the vertices of the dodecahedral cage has a diameter of around 2.4 nm (Fig. 3b)—only slightly larger than the diameter of primary silica clusters, being approximately 2 nm (ref. [18]; Extended Data Fig. 6). The interstitial spacing between two nearby vertices is estimated to be about 1.4 nm (this is the length of an edge, 3.8 nm, minus the diameter of the vertices, 2.4 nm; see Fig. 3b), which is much less than the diameter of such clusters. Bridges between vertices forming the edges of the dodecahedron are substantially thinner than the size of the primary clusters (Fig. 3a, b). This suggests that negatively charged primary silica clusters formed in solution may start to descend onto the positively charged micelle surface, attracted by Coulomb interactions. As more and more silica clusters assemble on the micelle surface, their repulsive interactions and possible interactions with other micelles may cause them to move to the vertices of a dodecahedron. Further silane condensation onto the surface of growing clusters may lead eventually to bridge formation, resulting in the final observed cage structure (Fig. 3). The origin of icosahedral symmetry in viruses has been associated with the energy minimization

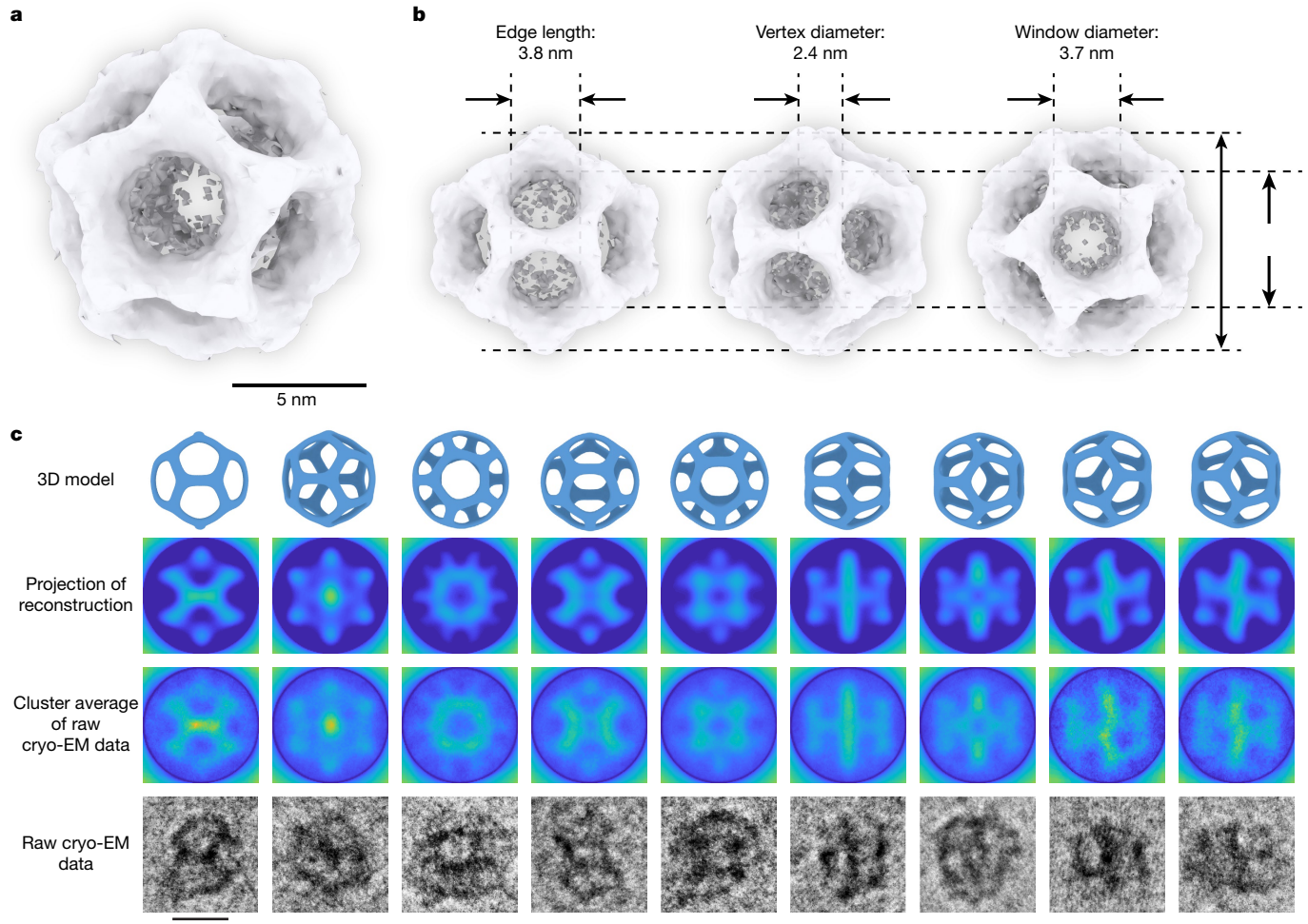


Fig. 3 | Single-particle reconstruction of the dodecahedral silicage. **a**, **b**, Reconstruction of the dodecahedral silicage (**a**), and its three most unique projections along the two-fold, three-fold and five-fold symmetry axes (**b**). The average dimensions of silicages in **b** were estimated from the reconstructed dodecahedral silicage. **c**, Representative comparison of nine unique projections from the reconstruction and cryo-EM cluster averages with projections of a 3D dodecahedral cage model. Corresponding single cryo-EM images are shown at the bottom, highlighting the difference between the raw data and the reconstruction. The scale bar in **c** represents 10 nm. The visualizations in **a** and **b** were made using UCSF Chimera[23].

of two opposing interactions: repulsive interactions associated with bending rigidity, and attractive hydrophobic interactions[27]. In a similar way, in addition to electrostatic interactions, deformation of the micelle surface around the silica clusters may be another important contributor to the free energy in our system. This idea is supported by experiments showing that the cage structures do not form in the absence of TMB (Extended Data Fig. 7), which is expected to enhance the deformability of the micelle surface.

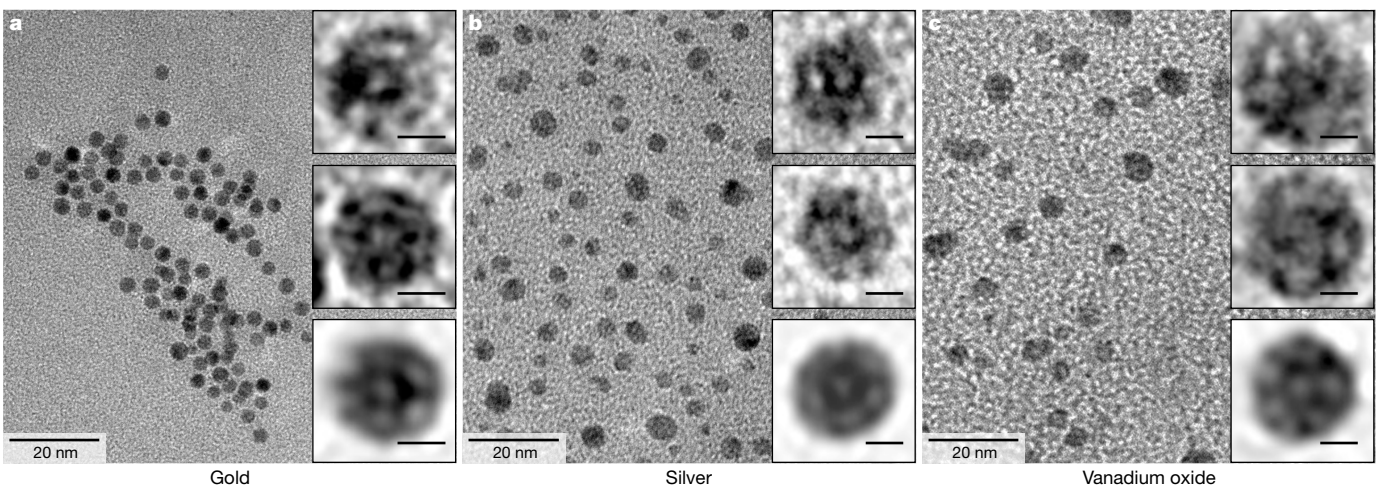


Fig. 4 | Cage-like structures with different inorganic compositions. **a**–**c**, Cage-like nanoparticles similar to silicages were obtained when silica was replaced by other materials, including gold (**a**), silver (**b**) or vanadium oxide (**c**). The insets show zoomed-in images of individual particles (top two rows) and averaged images[20] (bottom row). The scale bars in all insets represent 2 nm.

We have found that the self-assembly of ultrasmall cage structures directed by surfactant micelles can also occur when using other inorganic materials that have similarly sized features and similar surface chemistry to silica. In preliminary experiments, we replaced silica with one of two metals, gold or silver, or a transition metal oxide, vanadium oxide. Gold and silver structures were prepared by the reduction of metal precursors—gold chloride trihydrate and silver nitrate, respectively—in the presence of the micelles (Extended Data Fig. 8; see Methods). We used tetrakis(hydroxymethyl)phosphonium chloride (THPC) as both the reductant and the capping agent to stabilize primary gold and silver nanoparticles and to provide negative surface charges[28]. By contrast, primary vanadium oxide nanoparticles with a native negatively charged particle surface were prepared via sol–gel chemistry[29], in a similar way to the synthesis of silicages (see Methods). Images of individual particles obtained by TEM revealed similar internal structures (Fig. 4). These nanoparticles did not appear to be dense but instead showed cage-like structures (compare Figs. 2 and 4), as corroborated by associated projection averages[20] revealing cages with rotational symmetry (bottom insets in Fig. 4), similar to the prevalent projection in case of the silicage (vide supra). Therefore, micelle-self-assembly-directed cages like the dodecagonal structures described here may not be unique to amorphous silica, but may provide direct synthesis pathways to crystalline material cages (Extended Data Fig. 9).

There are several ramifications of our silicage discovery. First, such cages are generally considered to be individual structural units from which larger-scale mesoporous silica is built in a bottom-up manner[15,21]. However, given that a dodecahedron cannot be used to generate a tessellation of 3D space, other silica cage structures must be required. This knowledge should motivate better understanding of the early formation pathways of surfactant-directed silica self-assembly, including a search for micelle-directed ultrasmall silicages with other structures, and from other materials (vide supra). Second, the chemical and practical value of this polyhedral structure may prove immense. Given the versatility of silica surface chemistry, and the ability to distinguish the inside and outside of the cage via micelle-directed synthesis[17], one can readily conceive of cage derivatives of many kinds, which may exhibit unusual properties and be useful in applications ranging from catalysis to drug delivery. For example, given recent success in the clinical translation of ultrasmall fluorescent silica nanoparticles with similar particle sizes and surface properties to those described here[30], one can envisage a range of new diagnostic and therapeutic probes with drugs hidden inside the cages.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0221-0

Received: 17 December 2017; Accepted: 20 April 2018;
Published online: 20 June 2018

1. He, Y. et al. Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra. *Nature* **452**, 198–201 (2008).
2. Douglas, S. M. et al. Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* **459**, 414–418 (2009); erratum **459**, 1154 (2009).
3. Iinuma, R. et al. Polyhedra self-assembled from DNA tripods and characterized with 3D DNA-PAINT. *Science* **344**, 65–69 (2014).
4. Afonin, K. A. et al. *In vitro* assembly of cubic RNA-based scaffolds designed in silico. *Nat. Nanotechnol.* **5**, 676–682 (2010).
5. King, N. P. et al. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336**, 1171–1174 (2012).
6. Hsia, Y. et al. Design of a hyperstable 60-subunit protein icosahedron. *Nature* **535**, 136–139 (2016); corrigendum **540**, 150 (2016).
7. Rothemund, P. W. Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297–302 (2006).
8. Yin, P., Choi, H. M. T., Calvert, C. R. & Pierce, N. A. Programming biomolecular self-assembly pathways. *Nature* **451**, 318–322 (2008).
9. Ke, Y., Ong, L., Shih, W. & Yin, P. Three-dimensional structures self-assembled from DNA bricks. *Science* **338**, 1177–1183 (2012).
10. King, N. P. et al. Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103–108 (2014).
11. Zhao, M. et al. Mechanistic insights into the recycling machine of the SNARE complex. *Nature* **518**, 61–67 (2015).
12. Fernandez-Leiro, R. & Scheres, S. H. W. Unravelling biological macromolecules with cryo-electron microscopy. *Nature* **537**, 339–346 (2016).
13. Dai, X. et al. In situ structures of the genome and genome-delivery apparatus in a single-stranded RNA virus. *Nature* **541**, 112–116 (2017).
14. Kresge, C. T., Leonowicz, M. E., Roth, W. J., Vartuli, J. C. & Beck, J. S. Ordered mesoporous molecular sieves synthesized by a liquid-crystal template mechanism. *Nature* **359**, 710–712 (1992).
15. Sun, Y. et al. Formation pathways of mesoporous silica nanoparticles with dodecagonal tiling. *Nat. Commun.* **8**, 252 (2017).
16. Ma, K. et al. Control of ultrasmall sub-10 nm ligand-functionalized fluorescent core–shell silica nanoparticle growth in water. *Chem. Mater.* **27**, 4119–4133 (2015).
17. Ma, K., Sai, H. & Wiesner, U. Ultrasmall sub-10 nm near-infrared fluorescent mesoporous silica nanoparticles. *J. Am. Chem. Soc.* **134**, 13180–13183 (2012).
18. Carcouët, C. C. M. C. et al. Nucleation and growth of monodisperse silica nanoparticles. *Nano Lett.* **14**, 1433–1438 (2014).
19. Ma, K., Zhang, D., Cong, Y. & Wiesner, U. Elucidating the mechanism of silica nanoparticle PEGylation processes using fluorescence correlation spectroscopies. *Chem. Mater.* **28**, 1537–1545 (2016).
20. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
21. Xiao, C., Fujita, N., Miyasaka, K., Sakamoto, Y. & Terasaki, O. Dodecagonal tiling in mesoporous silica. *Nature* **487**, 349–353 (2012).
22. Gong, Y., Veesler, D., Doerschuk, P. C. & Johnson, J. E. Effect of the viral protease on the dynamics of bacteriophage HK97 maturation intermediates characterized by variance analysis of cryo EM particle ensembles. *J. Struct. Biol.* **193**, 188–195 (2016).
23. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
24. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
25. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from election micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
26. Harauz, G. & van Heel, M. Exact filters for general geometry three dimensional reconstruction. *Optik* **73**, 146–156 (1986).
27. Zandi, R., Reguera, D., Bruinsma, R. F., Gelbart, W. M. & Rudnick, J. Origin of icosahedral symmetry in viruses. *Proc. Natl Acad. Sci. USA* **101**, 15556–15560 (2004).
28. Duff, D. G., Baiker, A. & Edwards, P. P. New hydrosol of gold clusters. 1. Formation and particle size variation. *Langmuir* **9**, 2301–2309 (1993).
29. Sullivan, L. M., Li, L. & Lukehart, C. M. Synthesis of VO_2 nanopowders. Part I. Sol–gel processing of vanadium alkoxide precursor within inverse micelles. *J. Cluster Sci.* **25**, 313–322 (2014).
30. Phillips, E. et al. Clinical translation of an ultrasmall inorganic optical-PET imaging nanoparticle probe. *Sci. Transl. Med.* **6**, 260ra149 (2014).

Acknowledgements This project was supported by the National Cancer Institute of the National Institutes of Health under award number U54CA199081. Y.G. and P.C.D. acknowledge financial support from the National Science Foundation (NSF) under grant number 1217867, and Y.G. acknowledges financial support from a 2017 Google PhD Fellowship in Machine Learning. T.A. acknowledges financial support from the Ghent University Special Research Fund (BOF14/PDO/007) and the European Union's Horizon 2020 research and innovation program (MSCA-IF-2015-702300 and MSCA-RISE-691185). M.Z.T. acknowledges fellowship support from the Ministry of National Education of the Republic of Turkey. This work used shared facilities of the Cornell Center for Materials Research, with funding from the NSF Materials Research Science and Engineering Center program (DMR-1719875), as well as the Nanobiotechnology Center's shared research facilities at Cornell. The authors thank V. Elser, Y. Jiang and D. Zhang for helpful discussions.

Author contributions K.M., T.A. and U.W. designed the experimental work. Y.G. and P.C.D. performed the reconstructions. K.M. synthesized the silica-based materials. T.A. synthesized the metal-based and transition-metal-oxide-based materials. K.M. and T.A. performed TEM and cryo-EM characterization. K.M., T.A., M.Z.T. and T.K. processed the images for reconstructions. K.M., T.A. and U.W. discussed the experimental work. U.W. wrote the manuscript with input from all co-authors. U.W. supervised the work.

Competing interests The authors declare that they have submitted a patent disclosure based on this study through Cornell University.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0221-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0221-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to U.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Chemicals and materials. All chemicals were used as received. Cetyltrimethylammonium bromide (CTAB), ammonia (2 M in ethanol), mesitylene (1,3,5 trimethylbenzene; TMB), tetramethyl orthosilicate (TMOS), gold chloride trihydrate ($HAuCl_4{\cdot}3H_2O$), silver nitrate ($AgNO_3$), tetrakis(hydroxymethyl) phosphonium chloride (THPC), dimethyl sulfoxide (DMSO), acetic acid and ethanol were purchased from Sigma-Aldrich. Vanadium oxytriisopropoxide was purchased from Alfa Aesar. Anhydrous potassium carbonate (K_2CO_3) was purchased from Mallinckrodt. Anhydrous ethanol was purchased from Koptec. Silane-modified monofunctional polyethylene glycol (PEG-silane) with a molar mass of around 500 g mol^{-1} (six to nine ethylene glycol units) was purchased from Gelest. Carbon-film-coated copper grids for TEM, and C-Flat holey carbon grids for cryo-EM, were purchased from Electron Microscopy Sciences.

Synthesis, TEM and cryo-EM characterization of silicages. Silicages were synthesized in aqueous solution through surfactant-directed silica condensation. 125 mg of CTAB was first dissolved in 10 ml of ammonium hydroxide solution (0.002 M). Then, 100 μl TMB was added to expand the CTAB micelle size; the water/CTAB/TMB molar ratio was about 1,620/1/2. The solution was stirred at 600 r.p.m. at 30 °C overnight, and then 100 μl TMOS was added. The reaction was then left again at 30 °C overnight under stirring at 600 r.p.m.

To prepare cryo-EM samples, we applied 5 μl of the native reaction solution to a glow-discharged CF-4/2-2C Protochips C-Flat holey carbon grid, which was then blotted using filter paper and plunged into a liquid mixture of 37% ethane and 63% propane at −194 °C using an Electron Microscopy Science plunge freezer. Cryo-EM images were acquired on a FEI Tecnai F20-ST TEM operated at an acceleration voltage of 200 kV using a Gatan Orius charge-coupled-device (CCD) camera. All cryo-EM images used for reconstruction were acquired at the same magnification, with a pixel size of 0.16 nm, and nominal defocus was kept between 1 μm and 2 μm.

To prepare dry-state TEM samples, we added 100 μl of PEG-silane into the reaction solution. The reaction solution was left at 30 °C overnight under stirring at 600 r.p.m. to surface-modify silicages covalently with PEGs, in order to improve their dispersity on TEM grids. Afterwards, 30 μl of the reaction solution was dropped onto a copper grid coated with a continuous carbon film, and blotted using filter paper. TEM images were acquired using a FEI Tecnai T12 Spirit microscope operated at an acceleration voltage of 120 kV. In order to improve the signal-to-noise ratio in recorded images, TEM sample grids were plasma etched for 5 seconds before TEM characterization, and a series of images was acquired of the same sample area, which were then averaged.

In order to quench individual primary silica clusters formed at the very early stages of cage formation, we added 100 μl of PEG-silane into the reaction solution about three minutes after the addition of TMOS. The remainder of the procedures—including particle synthesis, dry-state TEM sample preparation, and TEM characterization—were as described above.

Particle purification. In order to remove CTAB and TMB from the cages, after adding PEG-silane and stirring at 30 °C for a day (see above), we heat-treated the solution at 80 °C overnight to further enhance the covalent attachment of PEG-silane to the silica surface of the silicages. The PEGylated nanocages were dialysed (with a molecular weight cut-off, MWCO, of 10 kDa) in a mixture of acetic acid, ethanol and water (volume ratio 7/500/500) for three days, and then in deionized water for another three days[17]. In both cases the dialysis solutions were changed once per day. The dry-state TEM sample preparation and TEM characterization methods were as described above.

Synthesis of particles without TMB. The synthesis and TEM characterization methods used for particles without TMB were the same as those for particles with TMB, as described above, except that the TMB addition step was omitted.

Silicage surface area and yield of production. We assessed the specific surface area of the silicages through a combination of nitrogen sorption measurements and theoretical estimations. After the synthesis and purification of PEGylated silicages, particles were first upconcentrated using a spin filter (Vivaspin 20, MWCO 10 kDa) and dried at 60 °C. Particles were then calcined at 550 °C for 6 hours in air. The production yield was then estimated by dividing the remaining weight after calcination, that is, the weight of inorganic silica, by the theoretical weight of silica, that is, the weight calculated on the basis of the amount of silica source added into the synthesis. Nitrogen adsorption and desorption isotherms were acquired using a Micromeritics ASAP 2020 (Extended Data Fig. 1c), yielding a specific surface area of 570 $m^2\ g^{-1}$ by the Brunauer–Emmett–Teller (BET) method. For comparison, using the dodecahedral cage model with the dimensions from the reconstruction shown in Fig. 3, we estimated the theoretical surface area of silicages to be around 790 $m^2\ g^{-1}$. The lower experimental value is consistent with the expected losses of surface area during sample calcination.

Synthesis and TEM of metal cage-like structures. The gold and silver cage-like structures were prepared by the reduction of metal precursors—$HAuCl_4{\cdot}3H_2O$ and $AgNO_3$, respectively—in the presence of micelles with the same water/CTAB/TMB ratio as for the silicage work. In a typical batch, 50 mg of CTAB was dissolved in 4 ml of water at 30 °C, then 40 μl of TMB and 200 μl of ethanol were added to the mixture. After stirring the reaction at 30 °C overnight at 600 r.p.m., 16 μl of either $HAuCl_4{\cdot}3H_2O$ (25 mM) or $AgNO_3$ (25 mM) was added, followed after 5 min by 8 μl of THPC (68 mM). After another 5 min, 6 μl of potassium carbonate (0.2 M) was added.

Dry-state TEM samples of gold and silver cage-like structures were prepared after one day and 6 hours of reaction, respectively, owing to the different reaction rates as described in further synthesis Methods sections below. In both cases, the samples were prepared by drying 8 μl of the native reaction mixture diluted three times in ethanol on a TEM grid in air overnight. In order to remove the thick CTAB layer before imaging, the grid was immersed in ethanol for 2 min and then dried in air. TEM images of metal cage-like structures were acquired using a FEI Tecnai T12 Spirit microscope operated at an acceleration voltage of 120 kV.

Synthesis and TEM of metal oxide structures. Vanadium oxide cage-like structures were prepared on the basis of sol–gel chemistry in a similar way to the silicages, using vanadium oxytriisopropoxide as the precursor. In a typical batch, 50 mg of CTAB was dissolved in 4 ml of water at 30 °C, and then 40 μl of TMB were added. After stirring the reaction at 30 °C overnight at 600 r.p.m., 50 μl of vanadium oxytriisopropoxide diluted in 100 μl of DMSO were added.

Dry-state TEM samples for vanadium oxide cage-like structures were prepared after one day of reaction by drying on a TEM grid 8 μl of the native reaction mixture, diluted ten times in water. At such a dilution, the amount of CTAB was low enough that the TEM samples did not require any plasma cleaning or soaking in ethanol before imaging. The TEM images of vanadium oxide cages were acquired using a FEI Tecnai T12 Spirit microscope operated at an acceleration voltage of 120 kV.

Particle reconstruction. We used the 'Hetero' model-based maximum likelihood machine-learning algorithm[22] for particle reconstruction. This algorithm can simultaneously estimate: (1) a reconstruction for each type of particle shown in the images; (2) the type of particle shown in each image; and (3) the projection orientation for each image. Such a joint estimation is a central feature of the algorithm and is a natural approach for processing data from complicated mixtures. The estimates in (2) and (3), which are based on 3D structure, are independent of the clustering of 2D images, which is based on pixel values (see, for example, Extended Data Fig. 2). In addition to the Hetero algorithm, we applied the widely used RELION 2.1 system[24] to compute equivalent two-class and single-class reconstructions. The images were corrected for the contrast transfer function (CTF) by phase flipping.

Further details of metal- and metal-oxide-based cages. In contrast to the sol–gel reaction that produces the silicage, synthesis of the gold and silver cage-like structures relies on reduction reactions. To this end, we used THPC because it reacts in water at basic pH to form trimethoxyphosphine, which can be both the reductant and the capping agent for the metal nanoparticles. THPC has been widely used to synthesize ultrasmall (less than 3 nm) and negatively charged phosphine-stabilized gold nanoparticles[28]. These nanoparticles are often used as seeds for the subsequent growth of continuous gold shells on the surface of aminated silica nanoparticles thanks to their high affinity and binding efficiency with amine groups[31,32]. Alcohol was added to the reaction mixture in order to mimic the conditions of the silicage synthesis, where methanol is formed upon hydrolysis of TMOS. Early-stage preliminary experiments showed that the resulting metal structures improved in size dispersity when using ethanol at a slightly higher concentration than that of the released methanol in the silicage synthesis. Gold and silver cage-like structure syntheses were performed at a much lower concentration (silver or gold concentration = 93.7 μM) than that of the silicages (silica precursor concentration = 65.9 mM). Attempts at synthesizing gold and silver cages at higher concentrations resulted in much larger nanoparticles with no apparent internal structure.

Gold-based synthesis. The addition of gold precursor to the reaction initially resulted in the formation of a pale yellow precipitate that turned into a clear (that is, non-turbid), darker-orange solution within a couple of minutes under stirring at 30 °C (see Extended Data Fig. 8 for a survey of the absorption characteristics at each step of the synthesis). Given that neither the precipitate nor the darker-orange colouration was observed in the absence of CTAB, we attribute these observations to some interaction between the gold chloride anions and the ammonium groups of the CTAB. After the addition of THPC, the solution turned colourless within a couple of minutes, indicating that gold(III) had been reduced to gold(I). The subsequent transformation of THPC into trimethoxyphosphine, following an increase in pH through the addition of potassium carbonate, happened within the first hour of reaction (see also the description of silver-based synthesis below). However, the reduction from gold(I) to gold(0) was rather slow, with the first hint of colouration appearing after 8 hours of reaction. After one day of reaction, the solution exhibited a brown colouration. This colouration is the signature of gold nanoparticles that are too small or not dense enough to exhibit a strong surface plasmon resonance,

as shown by the absorption profile in Extended Data Fig. 8, which shows only a faint feature around 510 nm.

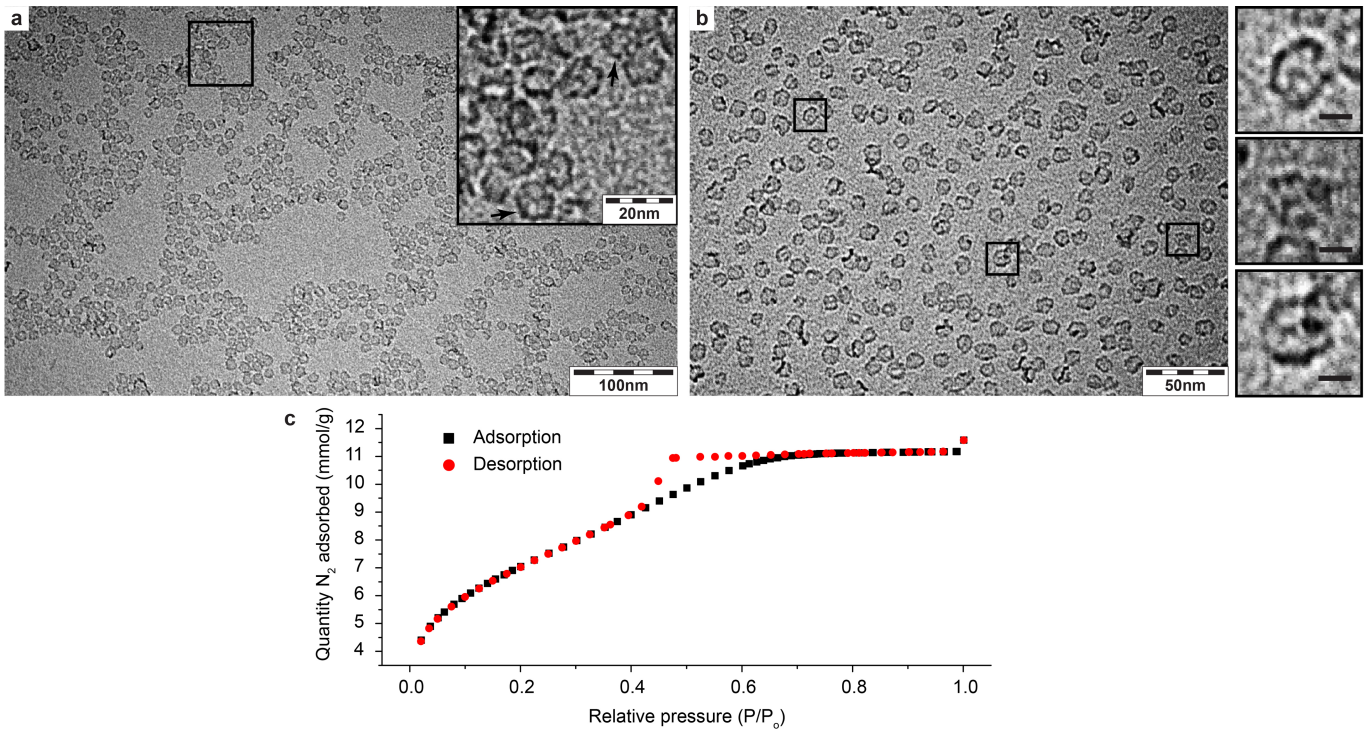
Silver-based synthesis. The addition of silver precursor to the reaction did not initially translate into any visible effects, either in the presence of CTAB/TMB or after adding THPC. Nevertheless, after adding potassium carbonate to the silver-based synthesis, the solution started to turn pale yellow within the first hour of reaction and resulted in an intense yellow colouration after 6 hours, at which point we prepared the TEM samples. This yellow colouration is typical of small silver nanoparticles with a surface plasmon resonance centred around 420 nm (Extended Data Fig. 8).

Vanadium-oxide-based synthesis. The vanadium oxide cage-like nanoparticles were prepared under the same conditions as the silicages; however, the pH was not adjusted with ammonia owing to the fast hydrolysis and condensation rate of the vanadium oxide precursor. Unlike in the synthesis of metal cage-like nanoparticles, no alcohol was added here, because the hydrolysis of the vanadium precursor—vanadium oxytriisopropoxide—produces alcohol at concentrations similar to those used in the silicage synthesis. The addition of this precursor to the TMB micelles resulted in the immediate formation of a red precipitate. Under stirring, the precipitate dispersed homogeneously in solution, which remained turbid, and turned orange after one day of reaction at 30 °C.

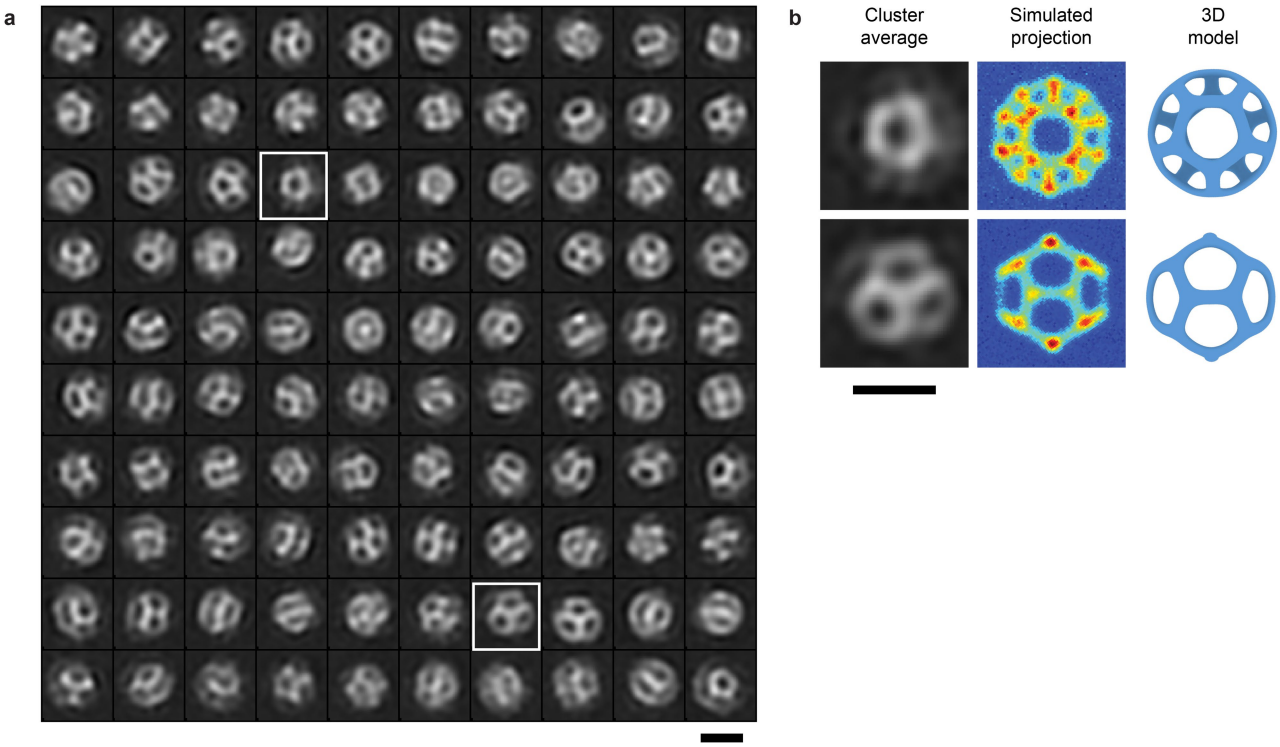
Data availability. Figures that have associated raw data are Figs. 2–4 and Extended Data Figs. 1–4, 5c and 6–9. There are no restrictions on data availability. The data sets generated and analysed during this study are available from the corresponding author on reasonable request.

Code availability. The custom code and algorithm used for the 3D reconstruction and related analysis are available from the corresponding author on reasonable request.

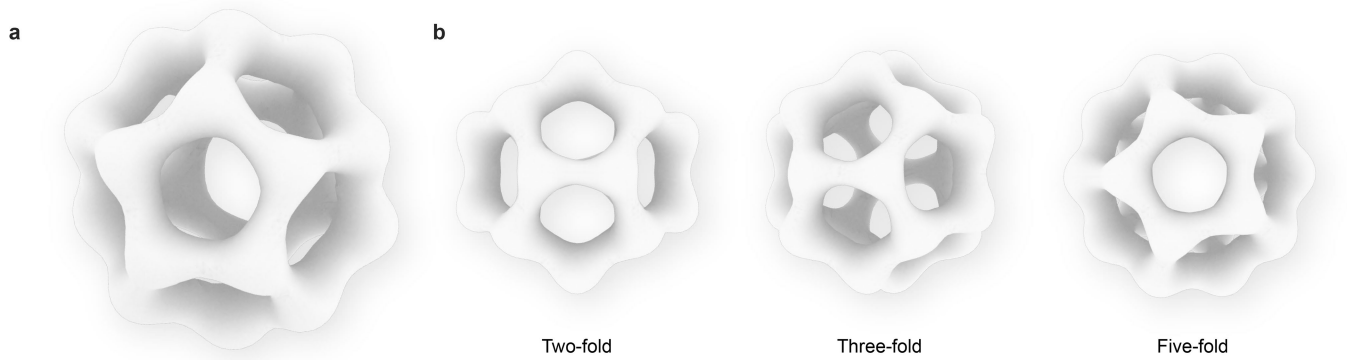
31. Westcott, S. L., Oldenburg, S. J., Lee, T. R. & Halas, N. J. Formation and adsorption of clusters of gold nanoparticles onto functionalized silica nanoparticle surfaces. *Langmuir* **14**, 5396–5401 (1998).
32. Ji, B. et al. Non-blinking quantum dot with a plasmonic nanoshell resonator. *Nat. Nanotechnol.* **10**, 170–175 (2015).



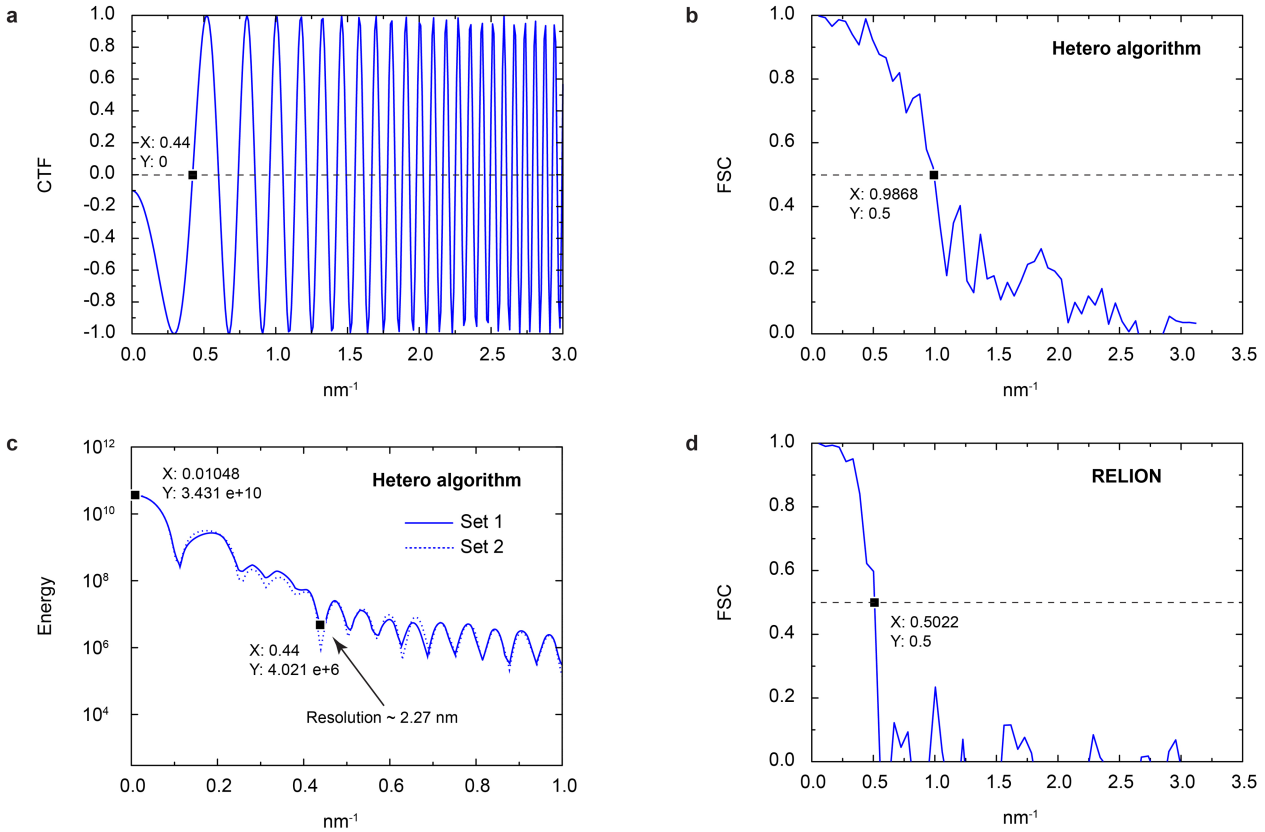
Extended Data Fig. 1 | PEGylated silicages after cleaning, and nitrogen sorption measurements on calcined cages. a, **b**, Representative dry-state TEM images, at different magnifications, of PEGylated silicages after the removal of surfactant (CTAB) and TMB (see Methods). The insets in **a** (black arrows) and **b** reveal cage-like structures, suggesting structure preservation after the removal of CTAB and TMB. **c**, Nitrogen adsorption and desorption isotherms of calcined silicages. After CTAB and TMB were removed, particles were calcined at 550 °C for 6 hours in air before nitrogen sorption measurements were taken. A particle synthesis yield of 67% was estimated from the weight of the calcined powder. The surface area of calcined silicages, as assessed by the Brunauer–Emmett–Teller (BET) method, was 570 $m^2\ g^{-1}$, consistent with theoretical estimations (Methods). Scale bars in the insets in **b** represent 5 nm.



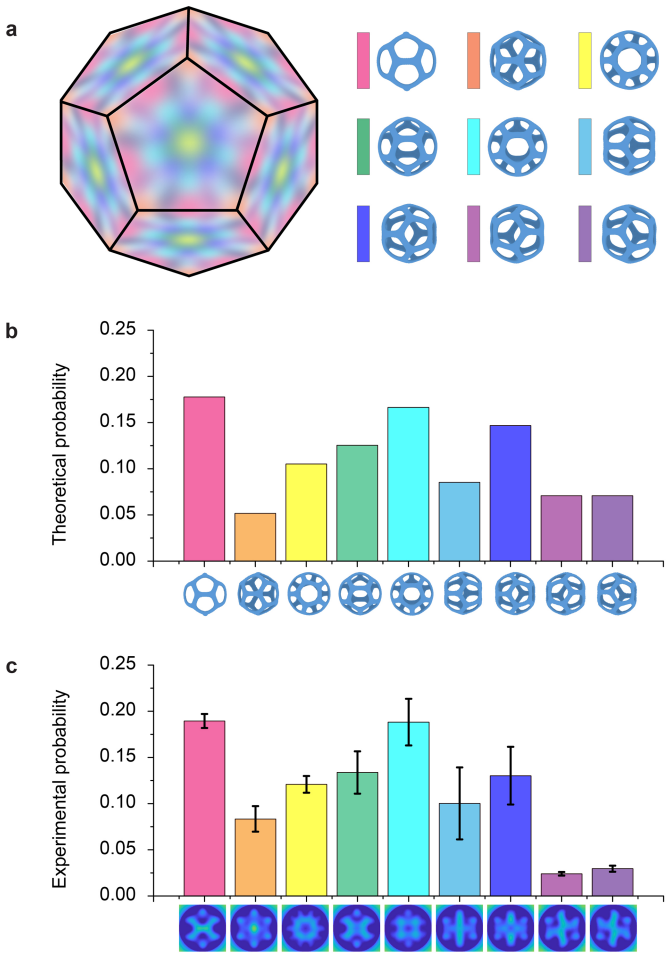
Extended Data Fig. 2 | Cluster averages of two-dimensional images of silicages. a, Around 19,000 single-particle cryo-EM images were sorted into 100 clusters[20]. **b**, Some of the projections (examples highlighted in **a**) exhibited features similar to those of simulated projections of dodecahedral cage structure. Also shown are projection models. Scale bars represent 10 nm.



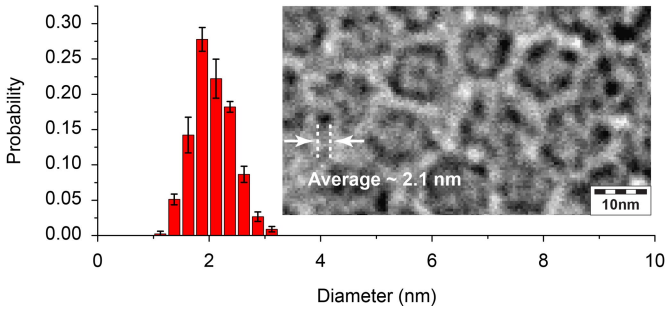
Extended Data Fig. 3 | Reconstruction of a silicage using the RELION 2.1 system. a, Reconstructed dodecahedral silicage[24]. **b**, Its three most unique projections, along the two-fold, three-fold and five-fold symmetry axes. The reconstruction was obtained from a single-class calculation run by RELION 2.1, using the same set of single-particle images as for the dodecahedral cage in Fig. 3a. Visualization is by UCSF Chimera[23].



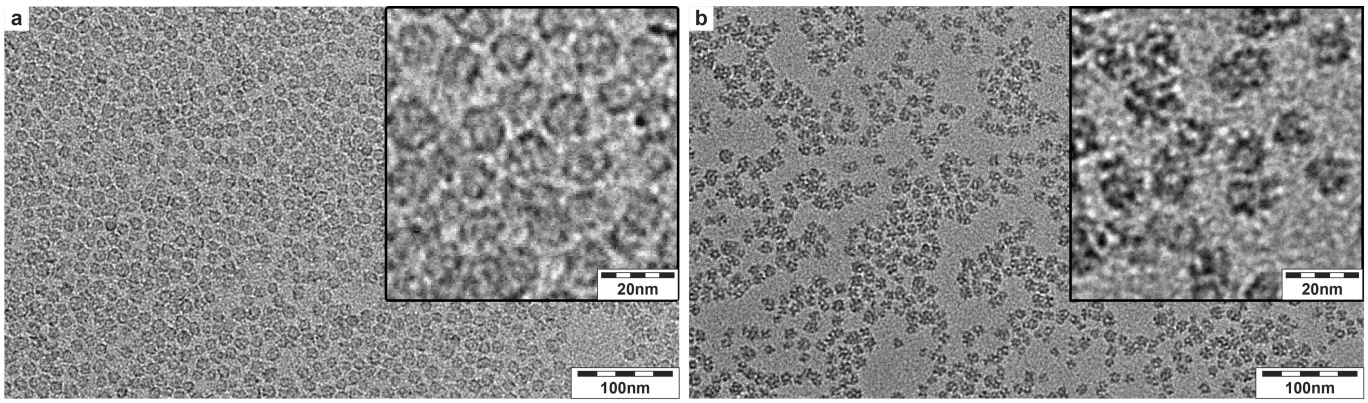
Extended Data Fig. 4 | A typical contrast transfer function, and determination of reconstruction resolution. We used CTFFIND4.1.8 (ref. [25]) to estimate defocus for individual micrographs or a set of micrographs, with results consistent with nominal defocus values of 1–2 μm. **a**, Contrast transfer function (CTF) for a defocus of 1.98 μm. Given that the first zero-crossing of the CTF occurs at 0.44 nm^{-1}, the CTF has little effect on reconstructions unless the resolution is greater than 1/(0.44 nm^{-1}), that is 2.27 nm. **b**, A Fourier shell correlation (FSC)[25] computed by a standard package[20] for two Hetero reconstructions that are independent, starting at the level of separate sets of images each containing 2,000 images (that is, 'gold standard' FSC). The resolution implied by the FSC curve (the inverse of the value of spatial frequency where the FSC curve first crosses 0.5) is 1/(0.99 nm^{-1}), that is 1.01 nm. **c**, Energy function for the same pair of reconstructions as in panel **b**. The energy is the spherical average of the squared magnitude of the reciprocal-space electron-scattering intensity, where the denominator of FSC is the square root of a product of two energy functions, one for each reconstruction. The observation that energy has dropped by more than 10^{-3} times its peak value, and that the character of the curve has become oscillatory and more slowly decreasing—both by 0.44 nm^{-1}—indicates that the resolution implied by the FSC curve is exaggerated[22] and that a more conservative resolution is 1/(0.44 nm^{-1}), that is 2.27 nm. **d**, FSC computed by a standard package[20] for two RELION 2.1 reconstructions computed from the same images as those in panel **b**, from which the resolution (at 0.5 threshold) is estimated to be around 1/(0.50 nm^{-1}), that is 2.00 nm.



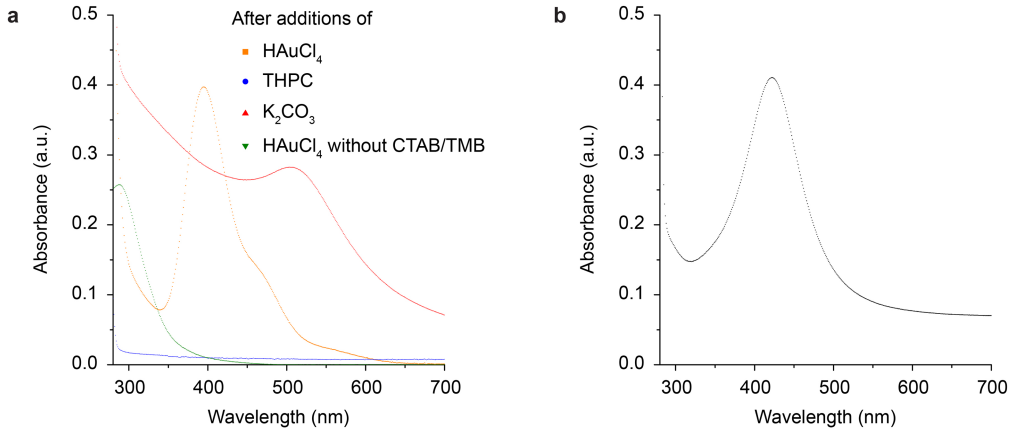
Extended Data Fig. 5 | Probability analysis of silicage projections. **a**, Orientation dependence of silicage projections. Right panel, the nine different silicage projections identified by 3D reconstruction (Fig. 3) are shown. Left panel, these orientations are manually mapped onto the surface of a dodecahedron. The orientations corresponding to different projections are assigned different colours. **b**, Probability analysis for different silicage projections. The probability of imaging a particular projection by electron microscopy is estimated by dividing that subset of the surface area of a sphere that contains the orientations corresponding to a specific projection, by the total surface area of the sphere. **c**, Experimental probability of imaging different silicage projections. The probability of imaging each projection is calculated by dividing the number of single-particle images assigned to a specific silicage projection via 3D reconstruction, by the overall number of silicage single-particle images. The error bars in **c** are standard deviations calculated from three projection distributions, which were obtained from three independent reconstruction runs using different sets of single-particle images.



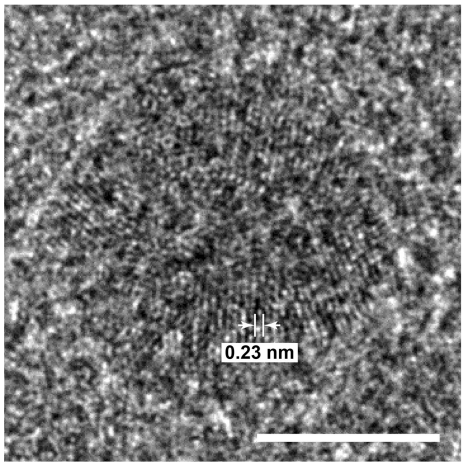
Extended Data Fig. 6 | Size analysis of silica clusters at an early stage of cage formation. Particle-size distribution for primary silica clusters at an early stage of cage formation, obtained by manually analysing 450 silica clusters using a set of TEM images. The measured silica clusters were randomly split into three groups, each containing 150 particles. A cluster-size distribution was then obtained for each of the three groups, and the results were averaged. The error bars are standard deviations calculated from the three cluster-size distributions. A representative TEM image is included at the right. In order to quench the very early stages of cage formation, PEG-silane was added into the synthesis mixture about three minutes after the addition of TMOS, thereby PEGylating early silica structures. TEM sample preparation and characterization were as described in the Methods. Primary silica clusters with diameters of around 2 nm were identified, consistent with the proposed cage-formation mechanism.



Extended Data Fig. 7 | Role of TMB in cage formation. a, **b**, TEM images, at different magnifications, of silica nanoparticles synthesized with (**a**) and without (**b**) TMB. Nanoparticles synthesized without TMB (**b**) show stronger contrast at the particle centres than do the nanocages (**a**), suggesting that these nanoparticles do not exhibit a hollow cage-like structure but instead are conventional mesoporous silica nanoparticles with relatively small particle sizes.



Extended Data Fig. 8 | Optical characterization of gold- and silver-based synthesis solutions. a, Survey of the gold-based synthesis, showing the absorption profile of solutions after the successive addition of $HAuCl_4$ (orange) and THPC (blue) and then one day after the addition of K_2CO_3 (red); also shown is the absorption profile when the same concentration of $HAuCl_4$ is added to the equivalent water/ethanol solution but without any CTAB or TMB (green). **b**, Absorption profile of a solution obtained from the silver synthesis 6 hours after the addition of K_2CO_3.



Extended Data Fig. 9 | High-resolution TEM image of a single cage-like gold nanoparticle. This gold particle exhibits lattice fringes with a spacing of 2.3 Å, consistent with the known lattice spacing between (111) planes of gold (Joint Committee on Powder Diffraction Standards no. 04-0784, http://www.icdd.com/). Scale bar represents 5 nm.

LETTER

https://doi.org/10.1038/s41586-018-0220-1

Enantioselective remote *meta*-C–H arylation and alkylation via a chiral transient mediator

Hang Shi[1], Alastair N. Herron[1], Ying Shao[1], Qian Shao[1] & Jin-Quan Yu[1]*

Enantioselective carbon–hydrogen (C–H) activation reactions by asymmetric metallation could provide new routes for the construction of chiral molecules[1,2]. However, current methods are typically limited to the formation of five- or six-membered metallacycles, thereby preventing the asymmetric functionalization of C–H bonds at positions remote to existing functional groups. Here we report enantioselective remote C–H activation using a catalytic amount of a chiral norbornene as a transient mediator, which relays initial *ortho*-C–H activation to the *meta* position. This was used in the enantioselective *meta*-C–H arylation of benzylamines, as well as the arylation and alkylation of homobenzylamines. The enantioselectivities obtained using the chiral transient mediator are comparable across different classes of substrates containing either neutral σ-donor or anionic coordinating groups. This relay strategy could provide an alternative means to remote chiral induction, one of the most challenging problems in asymmetric catalysis[3,4].

Enantioselective activation of C(sp^2)–H bonds has played an important role in the development of chiral ligands and the understanding of chiral induction in the metallation of C–H bonds. An early example of Ru(0)-catalysed atropselective alkylation of 2-arylpyridine[5] afforded only 49% enantiomeric excess (e.e.). Over the past decade, there has been an extensive search for suitable metal catalysts, chiral ligands and catalytic cycles that can achieve highly enantioselective C–H activation

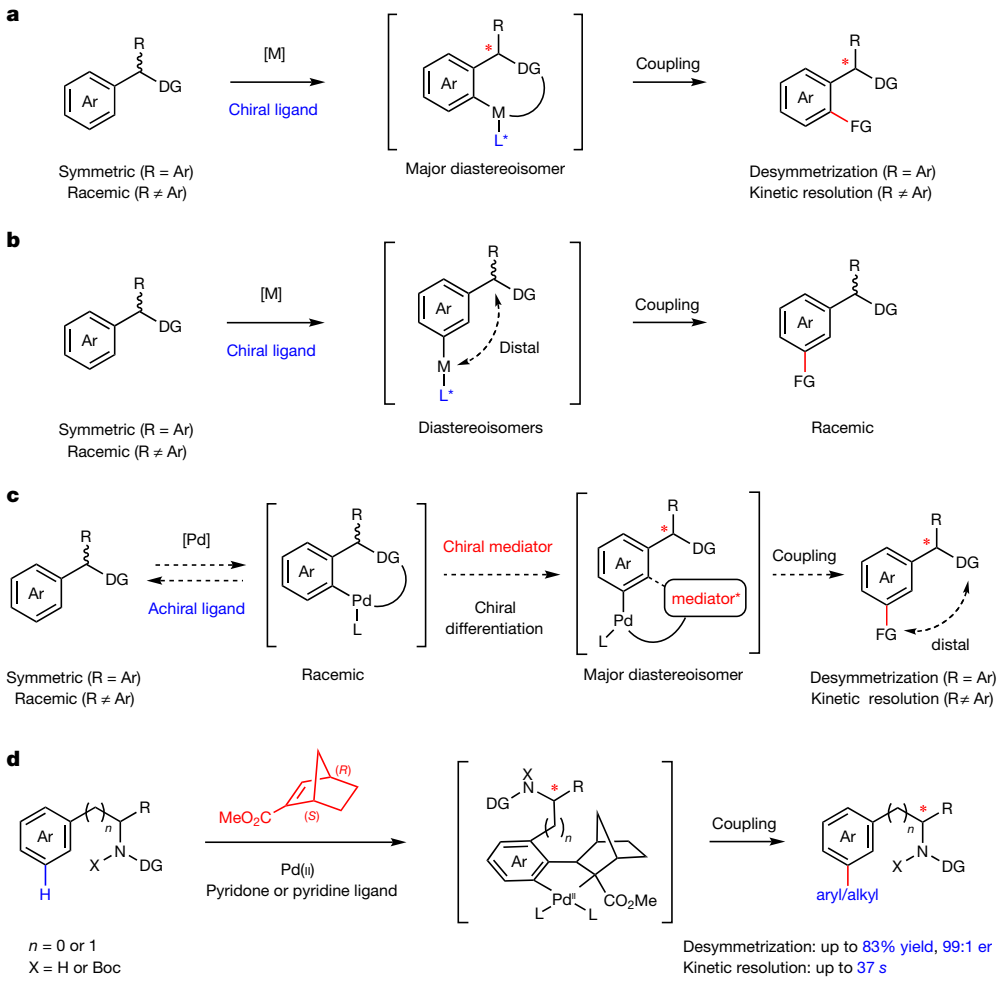
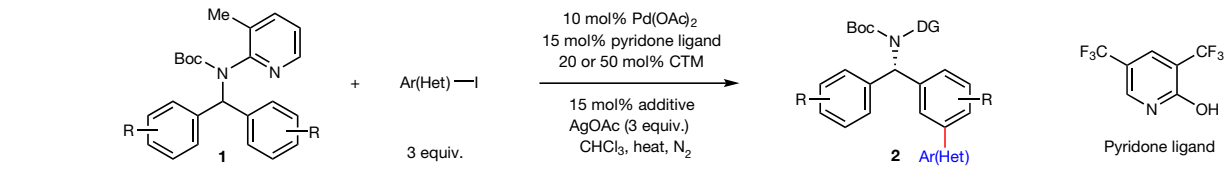
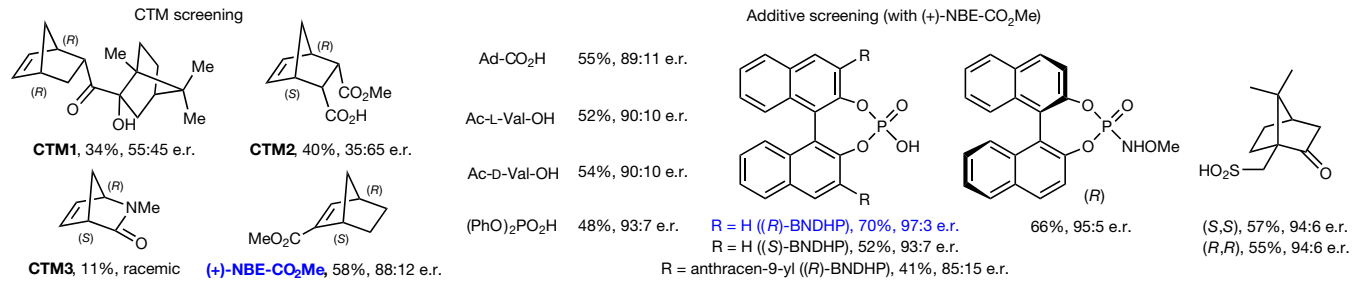


Fig. 1 | Enantioselective C(sp^2)–H activation. a, Enantioselective *ortho*-C–H activation. **b**, Enantioselective remote C–H activation, which is currently an unsolved problem. **c**, A strategy for enantioselective remote C–H activation. **d**, The scope of enantioselective remote C–H activation. Ar, aryl group; DG, directing group; FG, functional group; [M], transition-metal catalyst.

[1]Department of Chemistry, The Scripps Research Institute, La Jolla, CA, USA. *e-mail: yu200@scripps.edu



a



b

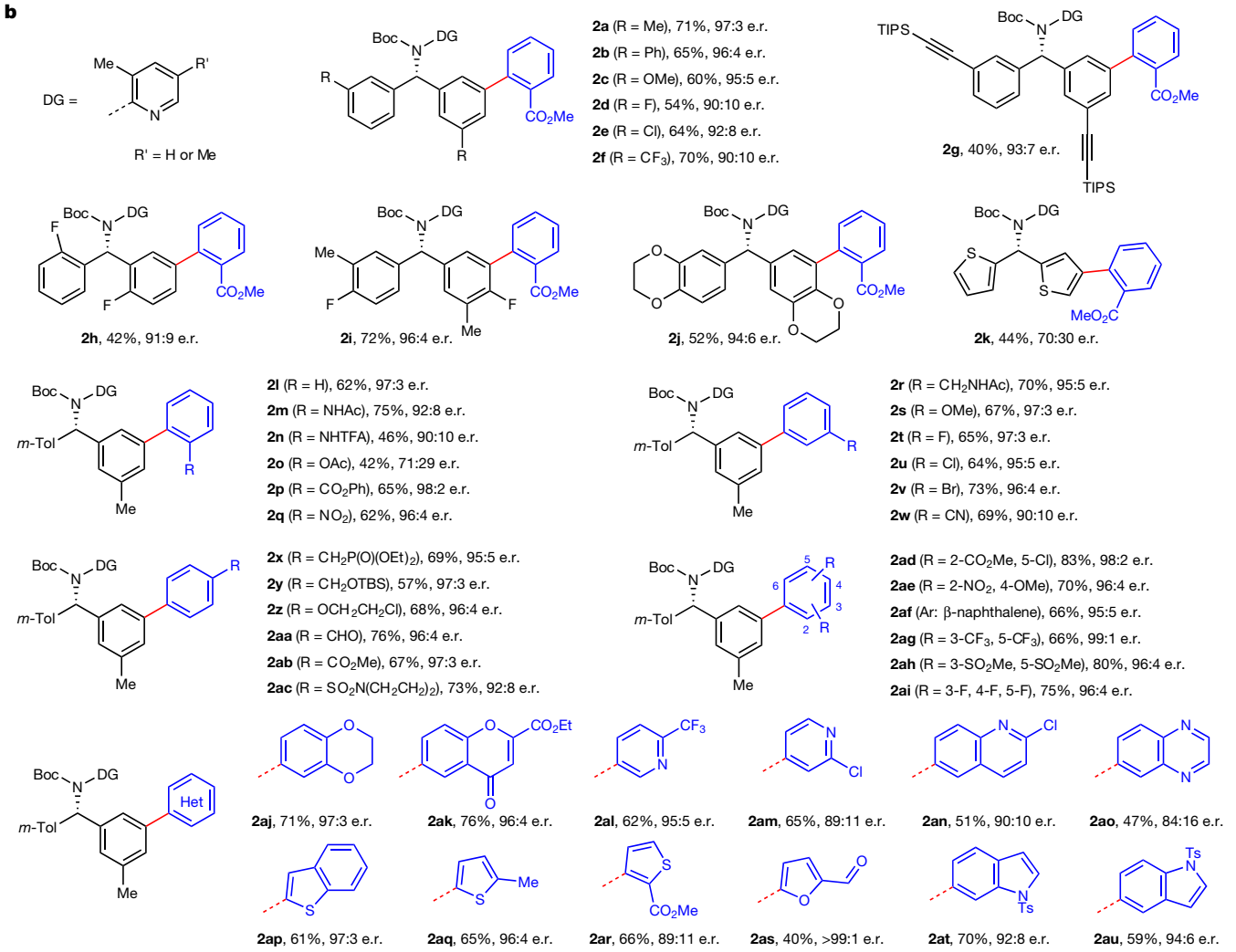


Fig. 2 | Enantioselective *meta*-C–H arylation of diarylmethylamines. **a**, CTM and additive optimization. Reaction conditions: 10 mol% $Pd(OAc)_2$, 15 mol% ligand, 50 mol% CTM, 15 mol% additive, 1 equiv. substrate (R = 3-Me), 3 equiv. methyl 4-iodobenzoate, 3 equiv. AgOAc, $CHCl_3$, 100 °C. Yields reported are based on NMR. **b**, The scope of asymmetric arylation. Isolated yields are reported. The exact reaction conditions for each substrate can be found in the Supporting Information. The absolute configuration of **2ah** was determined by X-ray crystallography. Reducing the $Pd(OAc)_2$ to 5 mol% produced comparable results for substrate **1a′** (see Supplementary Table 6). Ad-CO_2H, 1-adamantanecarboxylic acid; Ar(Het), (heterocyclic) aryl group; BNDHP, 1,1′-binaphthyl-2,2′-diyl hydrogen phosphate; Boc, *tert*-butyloxycarbonyl group; equiv., equivalents; TFA, trifluoroacetic acid; TIPS, triisopropylsilyl group; Tol, tolyl group.

reactions. This led to the finding that a Pd(II) catalyst bound to a chiral monoprotected amino acid ligand enables highly enantioselective C–H coupling via a Pd(II)/Pd(0) catalytic cycle[6–10], as well as the development of diverse transformations involving Pd(II)/Pd(IV) catalysis[11,12] (Fig. 1a). Recently, a number of enantioselective C–H activation reactions using chiral Pd(0)[13,14], Rh(I)[15,16], Rh(III)[17] and Ir(I)[18] catalysts have been reported. However, transition-metal-catalysed enantioselective remote *meta*-C(sp^2)–H activation has

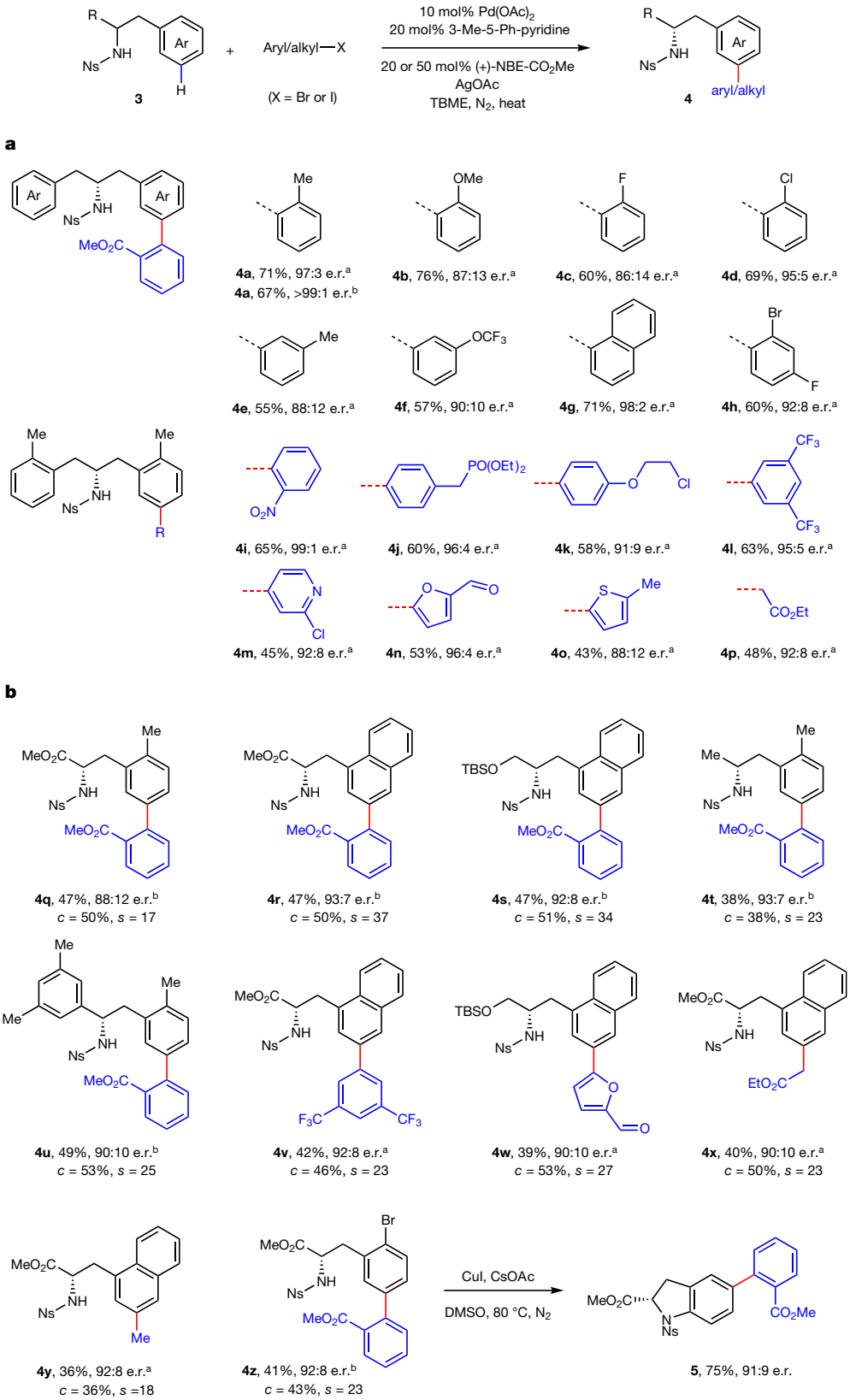


Fig. 3 | Enantioselective *meta*-C–H activation of homobenzylamines. **a**, The scope of desymmetrization. **b**, The scope of kinetic resolution. The exact reaction conditions for each substrate can be found in the Supporting Information. Isolated yields are reported. The absolute configurations of **4m** and **4n** were determined by X-ray crystallography. Reducing the catalyst loading to 5 mol% produced comparable results for substrate **3a** (see Supplementary Table 10). *c*, conversion; *s*, selectivity; TBME, *tert*-butylmethylether; TBS, *tert*-butyldimethylsilyl group. [a]Aryl or alkyl iodide as coupling reagent. [b]Methyl 2-bromobenzoate as coupling reagent.

hitherto not been achieved[19–22]. Asymmetric metallation of remote C(sp^2)–H bonds presents a distinct challenge: the metal insertion event occurs too far from the carbon that becomes the chiral centre. Remote C–H functionalization directed by a nitrile template has been developed recently[21]; however, control of the stereochemistry through a conformationally flexible macrocyclic transition state remains to

be demonstrated (Fig. 1b). To address this challenge, we devised a strategy using a catalytic amount of chiral transient mediator (CTM) to achieve enantioselective remote C–H activation (Fig. 1c). The CTM has a dual role: to relay C–H activation from the *ortho* position to the remote *meta* position[23–27]; and to achieve chiral differentiation of the racemic *ortho*-C–H palladation intermediates, generated by an achiral catalyst. The use of chiral norbornene methyl (1*S*,4*R*)-bicyclo[2.2.1]hept-2-ene-2-carboxylate ((+)-NBE-CO_2Me) as the CTM afforded high enantioselectivity in remote *meta*-C–H desymmetrization and kinetic resolution (Fig. 1d).

Our experimental design was based on a previous finding that norbornene can intercept the *ortho*-palladation intermediate via migratory insertion and subsequently effect *meta*-C–H functionalizations[23–25]. This reaction mechanism led us to consider that a chiral, enantioenriched norbornene might be able to differentiate the racemic *ortho*-palladate intermediate during the alkene insertion, or the subsequent alkene-insertion intermediate from the *meta*-C–H palladation step. We selected diarylmethylamine **1**[28], which contains a σ-donor coordinating group, as a model substrate to investigate the potential asymmetric arylation using chiral norbornene mediators (Fig. 2a). Although the chiral bicyclo[2.2.1]heptene derivatives **CTM1** and **CTM2** were effective mediators, both gave negligible enantioselectivity. Likewise, the lactam **CTM3** failed to display any degree of stereoinduction. The use of chiral (+)-NBE-CO_2Me as the transient mediator led to a marked improvement in enantioselectivity, giving an enantiomeric ratio (e.r.) of 88:12 (see Supplementary Tables 1–6 for optimization). Although neither inorganic base nor carboxylic acid additives were effective, the phosphoric acid $(PhO)_2PO_2H$ increased the enantiomeric ratio to 93:7. With (*R*)-BNDHP (1,1′-binaphthyl-2,2′-diyl hydrogen phosphate) as an additive, a 70% yield and an excellent enantioselectivity (97:3 e.r.) were obtained. Bulky substituents (such as an anthracenyl group) on the phosphoric acid proved detrimental to the reaction. To investigate the role of the chiral norbornene and phosphoric acid, we performed a number of control experiments. The use of racemic NBE-CO_2Me and (*R*)-BNDHP gave poor enantioselectivity (57:43 e.r.). The use of (*S*)-BNDHP, (*R*)-phosphorylimide or chiral camphorsulfonic acid in combination with (+)-NBE-CO_2Me led to a slight decrease in yield and enantio selectivity. These combined experiments suggest that (+)-NBE-CO_2Me is responsible for the chiral induction and that the chiral phosphoric acid has a minor beneficial effect.

Using the optimized conditions, we then tested the substrate scope with methyl 2-iodobenzoate (Fig. 2b). Electron-neutral (**2a**, **2b**), electron-donating (**2c**) and electron-withdrawing (**2d** to **2f**) substituents were all well tolerated, providing high enantioselectivities (up to 97:3 e.r.). Substrate **1g**, bearing an alkyne group, gave a lower yield owing to the instability of the C≡C triple bond under the standard conditions. Substituents at the *ortho* position (**2h**) reduced the reactivity, presumably because of a steric effect, but high enantioselectivity (91:9 e.r.) was preserved. Bis-substituted arenes **2i** and **2j**, which contain heterocycles, were also compatible and consistently high enantioselectivity was observed. The five-membered heterocycle **2k** gave lower selectivity, probably owing to a coordinative effect.

We next surveyed the scope of aryl iodides (Fig. 2b). The scope of *ortho*-substituted aryl iodides (**2m**–**2q**) is broad, providing high enantioselectivities (up to 97:3 e.r.). Electronically diverse *meta*-substituted aryl iodides (**2r**–**2w**) were tolerated, and produced enantiomeric ratios of greater than 95:5 (**2s**–**2v**). The scope of *para*-substituted aryl iodides is also broad, and electronic factors do not affect the enantioselectivity (**2x**–**2ac**). Multi-substituted aryl iodides (**2ad**–**2ai**) were also tolerated under the conditions, and the enantiomeric ratios all exceeded 95:5. Among the aryl iodides tolerated were those featuring halogens (for example, in **2t**–**2v**) and reactive groups (NHAc in **2m** and **2r**, phosphonate moiety in **2x**, alkyl chloride **2z** and aldehyde in **2aa**) that can serve as useful synthetic handles for subsequent chemical manipulation. Importantly, heterocyclic aryl iodides also proved successful in the reaction (**2aj**–**2au**). Pyridine, quinoline, quinazoline, furan, thiophene and indole derivatives were all suitable coupling partners, providing good to high enantioselectivities. The simple removal of the directing group was demonstrated with product **2ae** (see Supplementary Information).

We then investigated the asymmetric *meta*-C–H activation of homobenzylamines[29]. To our knowledge, no *ortho*-C–H activation method to date has achieved the construction of a homobenzylic chiral centre. Proposing that the challenge might be overcome by the CTM strategy, we tested different mediators and found that (+)-NBE-CO_2Me remained the most effective, giving an enantiomeric ratio of 97:3 with nosyl (Ns)-protected homobenzylamine **3a**, which contains an anionic coordinating group (see Supplementary Tables 7–10). Notably, the use of a nosyl-protected amino moiety as the directing group represents a practical advantage in synthetic applications. Further investigation revealed that this method is compatible with a broad range of symmetric homobenzylamines (Fig. 3a). *Ortho*-, *meta*-, and *para*-substituents were all well tolerated, and provided the desired products (**4a**–**4f**, **4h**) in good to high enantioselectivities. Moreover, both naphthalene (**4g**) and the bis-substituted arene (**4h**) were amenable to the standard conditions.

The reaction tolerated electronically diverse aryl iodides (**4i** to **4l**) and provided enantiomeric ratios of up to 99:1. Reactions involving heterocyclic aryl iodides such as pyridine, furan and thiophene (**4m** to **4o**) were also successful. In addition to aryl iodides, 2-bromobenzoate was also tested and product **4a** was obtained in moderate yield and excellent enantioselectivity (>99:1 e.r.). Similarly, the aliphatic coupling partner iodoacetate gave product **4p** with an enantiomeric ratio of 92:8.

In addition to desymmetrization, the kinetic resolution of secondary amines was also achieved (Fig. 3b). (+)-NBE-CO_2Me successfully resolved a racemic mixture of amino esters; with 2-bromobenzoate, *meta*-arylated products (**4q**, **4r**) were obtained in good enantioselectivity along with the recovery of starting materials. These *meta*-arylated phenylalanines are central motifs in a family of bioactive natural products[30]. Alkyl- and aryl-substituted homobenzylamines (**4s**–**4u**) were also compatible with this system. Moreover, simple arene (**4v**), heterocycle (**4w**) and alkyl groups (**4x**, **4y**) were all enantioselectively introduced at the *meta* position by this method. To demonstrate the synthetic utility of this reaction, a copper-catalysed intramolecular amination of **4z** yielded chiral indoline **5** with the 5 position substituted, a core structure of the chemotherapy medication vincristine.

In summary, remote enantioselective C–H activation reactions were realized by relaying *ortho*-C–H activation to remote *meta*-C–H activation using a chiral norbornene as the mediator. The chiral amplification is achieved by fast, reversible, non-asymmetric C–H activation followed by enantioselective norbornene insertion or *meta*-C–H activation. This approach is compatible with substrates containing either neutral σ-donor or anionic coordinating groups.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0220-1.

Received: 15 January 2018; Accepted: 27 March 2018;
Published online: 18 June 2018

1. Giri, R., Shi, B.-F., Engle, K. M., Maugel, N. & Yu, J.-Q. Transition metal-catalyzed C–H activation reactions: diastereoselectivity and enantioselectivity. *Chem. Soc. Rev.* **38**, 3242–3272 (2009).
2. Newton, C. G., Wang, S.-G., Oliveira, C. C. & Cramer, N. Catalytic enantioselective transformations involving C–H bond cleavage by transition-metal complexes. *Chem. Rev.* **117**, 8908–8976 (2017).
3. Clayden, J., Lund, A., Vallverdú, L. & Helliwell, M. Ultra-remote stereocontrol by conformational communication of information along a carbon chain. *Nature* **431**, 966–971 (2004).
4. Hurtley, A. E., Stone, E. A., Metrano, A. J. & Miller, S. J. Desymmetrization of diarylmethylamido bis(phenols) through peptide-catalyzed bromination: enantiodivergence as a consequence of a 2 amu alteration at an achiral residue within the catalyst. *J. Org. Chem.* **82**, 11326–11336 (2017).

5. Kakiuchi, F., Gendre, P. L., Yamada, A., Ohtaki, H. & Murai, S. Atropselective alkylation of biaryl compounds by means of transition metal-catalyzed C–H/olefin coupling. *Tetrahedron: Asymmetry* **11**, 2647–2651 (2000).
6. Shi, B.-F., Maugel, N., Zhang, Y.-H. & Yu, J.-Q. Pd^{II}-catalyzed enantioselective activation of $C(sp^2)$–H and $C(sp^3)$–H bonds using monoprotected amino acids as chiral ligands. *Angew. Chem. Int. Ed.* **47**, 4882–4886 (2008).
7. Shi, B.-F., Zhang, Y. H., Lam, J. K., Wang, D. H. & Yu, J.-Q. Pd(II)-catalyzed enantioselective C–H olefination of diphenylacetic acids. *J. Am. Chem. Soc.* **132**, 460–461 (2010).
8. Du, Z. J. et al. Pd(II)-catalyzed enantioselective synthesis of P-stereogenic phosphinamides via desymmetric C–H arylation. *J. Am. Chem. Soc.* **137**, 632–635 (2015).
9. Gao, D.-W., Shi, Y.-C., Gu, Q., Zhao, Z.-L. & You, S.-L. Enantioselective synthesis of planar chiral ferrocenes via palladium-catalyzed direct coupling with arylboronic acids. *J. Am. Chem. Soc.* **135**, 86–89 (2013).
10. Pi, C. et al. Redox of ferrocene controlled asymmetric dehydrogenative Heck reaction *via* palladium-catalyzed dual C–H bond activation. *Chem. Sci.* **4**, 2675–2679 (2013).
11. Chu, L., Xiao, K.-J. & Yu, J.-Q. Room-temperature enantioselective C–H iodination via kinetic resolution. *Science* **346**, 451–455 (2014).
12. Gao, D.-W., Gu, Q. & You, S.-L. Pd(II)-catalyzed intermolecular direct C–H bond iodination: an efficient approach toward the synthesis of axially chiral compounds via kinetic resolution. *ACS Catal.* **4**, 2741–2745 (2014).
13. Albicker, M. R. & Cramer, N. Enantioselective palladium-catalyzed direct arylations at ambient temperature: access to indanes with quaternary stereocenters. *Angew. Chem. Int. Ed.* **48**, 9139–9142 (2009).
14. Shintani, R., Otomo, H., Ota, K. & Hayashi, T. Palladium-catalyzed asymmetric synthesis of silicon-stereogenic dibenzosiloles via enantioselective C–H bond functionalization. *J. Am. Chem. Soc.* **134**, 7305–7308 (2012).
15. Kuninobu, Y., Yamauchi, K., Tamura, N., Seiki, T. & Takai, K. Rhodium-catalyzed asymmetric synthesis of spirosilabifluorene derivatives. *Angew. Chem. Int. Ed.* **52**, 1520–1522 (2013).
16. Lee, T., Wilson, T. W., Berg, R., Ryberg, P. & Hartwig, J. F. Rhodium-catalyzed enantioselective silylation of arene C–H bonds: desymmetrization of diarylmethanols. *J. Am. Chem. Soc.* **137**, 6742–6745 (2015).
17. Sun, Y. & Cramer, N. Rhodium(III)-catalyzed enantiotopic C–H activation enables access to *P*-chiral cyclic phosphinamides. *Angew. Chem. Int. Ed.* **56**, 364–367 (2017).
18. Shibata, T. & Shizuno, T. Iridium-catalyzed enantioselective C–H alkylation of ferrocenes with alkenes using chiral diene ligands. *Angew. Chem. Int. Ed.* **53**, 5410–5413 (2014).
19. Saidi, O. et al. Ruthenium-catalyzed meta sulfonation of 2-phenylpyridines. *J. Am. Chem. Soc.* **133**, 19298–19301 (2011).
20. Hofmann, N. & Ackermann, L. meta-Selective C–H bond alkylation with secondary alkyl halides. *J. Am. Chem. Soc.* **135**, 5877–5884 (2013).
21. Leow, D., Li, G., Mei, T.-S. & Yu, J.-Q. Activation of remote *meta*-C–H bonds assisted by an end-on template. *Nature* **486**, 518–522 (2012).
22. Phipps, R. J. & Gaunt, M. J. A meta-selective copper-catalyzed C–H bond arylation. *Science* **323**, 1593–1597 (2009).
23. Wang, X.-C. et al. Ligand-enabled *meta*-C–H activation using a transient mediator. *Nature* **519**, 334–338 (2015).
24. Dong, Z., Wang, J. & Dong, G. Simple amine-directed meta-selective C–H arylation via Pd/norbornene catalysis. *J. Am. Chem. Soc.* **137**, 5887–5890 (2015).
25. Shen, P.-X., Wang, X.-C., Wang, P., Zhu, R.-Y. & Yu, J.-Q. Ligand-enabled *meta*-C–H alkylation and arylation using a modified norbornene. *J. Am. Chem. Soc.* **137**, 11574–11577 (2015).
26. Ye, J. & Lautens, M. Palladium-catalysed norbornene-mediated C–H functionalization of arenes. *Nat. Chem.* **7**, 863–870 (2015).
27. Della Ca', N., Fontana, M., Motti, E. & Catellani, M. Pd/norbornene: a winning combination for selective aromatic functionalization via C−H bond activation. *Acc. Chem. Res.* **49**, 1389–1400 (2016).
28. Wang, P., Farmer, M. E. & Yu, J.-Q. Ligand-promoted *meta*-C-H functionalization of benzylamines. *Angew. Chem. Int. Ed.* **56**, 5125–5129 (2017).
29. Ding, Q. et al. Ligand-enabled *meta*-selective C–H arylation of nosyl-protected phenethylamines, benzylamines, and 2-aryl anilines. *J. Am. Chem. Soc.* **139**, 417–425 (2017).
30. Albrecht, B. K. & Williams, R. M. A concise, total synthesis of the TMC-95A/B proteasome inhibitors. *Proc. Natl Acad. Sci. U.S.A.* **101**, 11949–11954 (2004).

Acknowledgements We acknowledge The Scripps Research Institute, the National Institutes of Health (National Institute of General Medical Sciences grant 5R01GM102265) and Shanghai RAAS Blood Products Co. Ltd for their financial support. Y.S. thanks Jiangsu Overseas Research & Training Program for University Prominent Young & Middle-aged Teachers and Presidents.

Reviewer information *Nature* thanks M. Catellani, G. Chen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.-Q.Y. and H.S. conceived the concept. H.S. developed the enantioselective remote C−H activation. H.S. and A.N.H. performed the mechanistic study. H.S., A.N.H., Y.S. and Q.S. prepared reaction substrates. J.-Q.Y. directed the project.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0220-1.

Correspondence and requests for materials should be addressed to J.-Q.Y.

METHODS

Mechanistic studies. Our previous mechanistic studies of enantioselective C–H activation reactions established the asymmetric C–H metallation step as the enantiodetermining step[31,32]. Because the chiral mediator (+)-NBE-CO_2Me is not involved in the *ortho*-C–H activation step, chiral differentiation should occur in one of the subsequent steps. A number of mechanistic experiments were therefore carried out to provide further insights (see Supplementary Figs. 1–8). First, kinetic isotope effects of 1.03 and 1.33 for the *ortho*- and *meta*-deuterated substrates, respectively, indicated that the C–H activation steps are probably not rate-limiting. The first-order rate dependence observed for aryl iodide points to the oxidative addition of aryl iodide as rate-limiting (see Supplementary Fig. 7)[33]. Second, substantial hydrogen–deuterium exchange at the *ortho* position and the lack of such exchange at the *meta* position suggests that the *ortho*-C–H activation step is fast and reversible whereas the *meta*-C–H activation is not readily reversible on the timescale of the reaction. These combined data suggest that chiral differentiation by the chiral norbornene occurs either during the norbornene insertion into the *ortho*-palladation intermediate or during the subsequent *meta*-C–H activation step (see Supplementary Fig. 9).

General procedure for the Pd/(+)-NBE-CO_2Me catalysed enantioselective *meta*-C–H activation. Arene (0.10 mmol, 1.0 equiv), $Pd(OAc)_2$ (2.2 mg, 10 μmol, 10 mol%), ligand (15 mol% or 20 mol%), hydrogen phosphate (15 mol% for benzylamine substrate), aryl iodide (0.3 mmol, 3.0 equiv.) and silver acetate (50 mg, 0.30 mmol, 3.0 equiv.) were added into a 2-dram reaction vial. Solvent and (+)-NBE-CO_2Me (20 mol% or 50 mol%) were added to the mixture. The vial was flushed with N_2 and capped. The reaction mixture was then stirred at the selected temperature for 12–24 h. After cooling to room temperature, the mixture was filtered through Celite and eluted with ethyl acetate. The filtrate was evaporated under reduced pressure. Purification by preparative thin-layer chromatography afforded the desired product. Full experimental details and characterization of new compounds can be found in the Supplementary Information.

Data availability. The data supporting the findings of this study are available within the article and its Supplementary Information Files. Metrical parameters for the structures of **2ah**, **4m** and **4n** are available free of charge from the Cambridge Crystallographic Data Centre (https://www.ccdc.cam.ac.uk/) under reference numbers 1586807, 1586808 and 1586809, respectively.

31. Musaev, D. G., Kaledin, A., Shi, B.-F. & Yu, J.-Q. Key mechanistic features of enantioselective C–H bond activation reactions catalyzed by [(chiral mono-*N*-protected amino acid)-Pd(II)] complexes. *J. Am. Chem. Soc.* **134**, 1690–1698 (2012).
32. Plata, R. E. et al. A role for Pd(IV) in catalytic enantioselective C–H functionalization with monoprotected amino acid ligands under mild conditions. *J. Am. Chem. Soc.* **139**, 9238–9245 (2017).
33. Jiao, L., Herdtweck, E. & Bach, T. Pd(II)-catalyzed regioselective 2-alkylation of indoles *via* a norbornene-mediated C–H activation: mechanism and applications. *J. Am. Chem. Soc.* **134**, 14563–14572 (2012).

LETTER

https://doi.org/10.1038/s41586-018-0222-z

Evidence for extremely rapid magma ocean crystallization and crust formation on Mars

Laura C. Bouvier[1,8], Maria M. Costa[1,8], James N. Connelly[1], Ninna K. Jensen[1], Daniel Wielandt[2], Michael Storey[2], Alexander A. Nemchin[3], Martin J. Whitehouse[4], Joshua F. Snape[4], Jeremy J. Bellucci[4], Frédéric Moynier[5], Arnaud Agranier[6], Bleuenn Gueguen[6], Maria Schönbächler[7] & Martin Bizzarro[1]*

The formation of a primordial crust is a critical step in the evolution of terrestrial planets but the timing of this process is poorly understood. The mineral zircon is a powerful tool for constraining crust formation because it can be accurately dated with the uranium-to-lead (U–Pb) isotopic decay system and is resistant to subsequent alteration. Moreover, given the high concentration of hafnium in zircon, the lutetium-to-hafnium (^{176}Lu–^{176}Hf) isotopic decay system can be used to determine the nature and formation timescale of its source reservoir[1–3]. Ancient igneous zircons with crystallization ages of around 4,430 million years (Myr) have been reported in Martian meteorites that are believed to represent regolith breccias from the southern highlands of Mars[4,5]. These zircons are present in evolved lithologies interpreted to reflect re-melted primary Martian crust[4], thereby potentially providing insight into early crustal evolution on Mars. Here, we report concomitant high-precision U–Pb ages and Hf-isotope compositions of ancient zircons from the NWA 7034 Martian regolith breccia. Seven zircons with mostly concordant U–Pb ages define ^{207}Pb/^{206}Pb dates ranging from 4,476.3 ± 0.9 Myr ago to 4,429.7 ± 1.0 Myr ago, including the oldest directly dated material from Mars. All zircons record unradiogenic initial Hf-isotope compositions inherited from an enriched, andesitic-like crust extracted from a primitive mantle no later than 4,547 Myr ago. Thus, a primordial crust existed on Mars by this time and survived for around 100 Myr before it was reworked, possibly by impacts[4,5], to produce magmas from which the zircons crystallized. Given that formation of a stable primordial crust is the end product of planetary differentiation, our data require that the accretion, core formation and magma ocean crystallization on Mars were completed less than 20 Myr after the formation of the Solar System. These timescales support models that suggest extremely rapid magma ocean crystallization leading to a gravitationally unstable stratified mantle, which subsequently overturns, resulting in decompression melting of rising cumulates and production of a primordial basaltic to andesitic crust[6,7].

The emergence of a stable primordial crust is a fundamental step in the early history of rocky, potentially habitable planets. Primordial crust formation is the end product of a long sequence of events, including planetary accretion, establishment of a global magma ocean, core formation and, finally, silicate differentiation. Existing constraints[6–12] for the timing of each of these events allow for primordial crust formation in terrestrial planets over timescales of approximately 5 to 100 Myr, a range that precludes a full understanding of early planet formation. In the Solar System, Mars offers the opportunity to better constrain the timing of planet-formation processes, given its relatively simple geologic history as a stranded planetary embryo[8] as well as the wealth of information from Martian meteorites and spacecraft exploration[13]. From the meteorite record, the accretion of Mars is inferred to have been mostly completed within about 5 Myr of the formation of the Solar System[8,14], whereas the crystallization of the magma ocean leading to the extraction of a primordial crust may have occurred over timescales of approximately 30 to 100 Myr after accretion[10,15,16]. However, these timescales for silicate differentiation are based on the modelled abundances of the short-lived ^{182}Hf and ^{146}Sm nuclides during planetary differentiation inferred from young Martian meteorites and, hence, are highly model dependent.

A more robust approach to dating early planetary differentiation on Mars requires the identification of material that formed in the earliest evolutionary stages of the planet. On Earth, such a record is preserved in the Jack Hills of Western Australia, which contains ancient zircons[17] as old as about 4,370 Myr. Although zircon is not common in Martian meteorites, two recent studies have reported the presence of approximately 4,430-Myr-old zircons in the Martian regolith breccia NWA 7533/7034, which is thought to have originated from the southern highlands of Mars[4,5]. The breccia contains clasts that are interpreted to be of igneous, sedimentary and impact origin, preserved in a fine-grained matrix. Zircons have been identified in the igneous and sedimentary clasts as well as within the matrix. Collectively, these grains are likely to provide the earliest tangible record of crust formation processes on Mars. However, the typical (small) sizes of these zircons prevent us from obtaining the concomitant U–Pb ages and Hf isotope compositions of sufficient precision using in situ techniques.

We conducted a systematic search for zircons in a bulk crushed rock aliquot of the NWA 7034 meteorite. Although this approach does not provide a petrological context for individual zircons, it is the only means of ensuring the recovery of grains sufficiently large (>30 μm) to permit high-precision U–Pb chronology and Hf isotope measurements using solution-based methods. Irrespective of their petrological context, these zircons faithfully record information about the nature of their source reservoir. This rationale is recognized in studies using the Jack Hills detrital zircons to probe the early terrestrial crustal record[17]. A total of seven grains were extracted and analysed for U–Pb dating and Lu–Hf systematics. Their sizes ranged from around 50 μm to 110 μm and they were found to represent different morphologic types, including rounded pieces, irregular anhedral pieces, and euhedral pieces with well-defined faces and a flat prismatic shape (Extended Data Fig. 1). Common to all of them was the general absence of fractures and inclusions, as well as any evidence of radiation damage. The zircons returned ^{207}Pb/^{206}Pb ages ranging from 4,476.3 ± 0.9 Myr to 4,429.7 ± 1.0 Myr (Table 1). Importantly, five of the seven grains are concordant within their stated uncertainties, grain S22b5 is 1.2% discordant and grain S24B2 is 5.3% discordant (Fig. 1). The larger degree of discordance for grain S24B2 is consistent with its higher U content of 46 parts per million (p.p.m.) relative to the other zircons we investigated. On the basis of textural information and geological context, the zircons from

[1]Centre for Star and Planet Formation and Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. [2]Quadlab and Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. [3]Department of Applied Geology, Curtin University, Perth, Western Australia, Australia. [4]Swedish Museum of Natural History, Stockholm, Sweden. [5]Institut de Physique du Globe de Paris, Université Paris Diderot, Sorbonne Paris Cité, Paris, France. [6]Laboratoire Géosciences Océan (UMR CNRS 6538), Université de Bretagne Occidentale et Institut Universitaire Européen de la Mer, Plouzané, France. [7]Institute of Geochemistry and Petrology, ETH, Zurich, Switzerland. [8]These authors contributed equally: Laura C. Bouvier, Maria M. Costa.
*e-mail: bizzarro@snm.ku.dk

Table 1 | U–Pb age data and ^{176}Lu–^{176}Hf systematics of NWA 7034 zircons and the Hf isotope composition of the 91500 terrestrial zircon standard

Sample	$^{207}Pb/^{206}Pb$ age (Myr)	$^{207}Pb/^{235}U$ age (Myr)	$^{206}Pb/^{238}U$ age (Myr)	$^{176}Lu/^{177}Hf$	$^{176}Hf/^{177}Hf$	$^{178}Hf/^{177}Hf$	$^{180}Hf/^{177}Hf$	εHf_T
S22b4	4,448.7 ± 1.8	4,445.1 ± 6.3	4,437.3 ± 22.1	0.000805	0.279891 ± 10	1.46718 ± 2	1.88666 ± 6	−1.92 ± 0.37
S22b5	4,474.2 ± 1.4	4,457.8 ± 5.7	4,421.9 ± 19.8	0.000799	0.279907 ± 09	1.46719 ± 2	1.88666 ± 5	−0.71 ± 0.32
S23b9	4,447.0 ± 1.5	4,441.0 ± 4.8	4,427.7 ± 16.5	0.000911	0.279908 ± 16	1.46721 ± 2	1.88669 ± 6	−1.70 ± 0.58
S24b2	4,429.7 ± 1.0	4,355.2 ± 4.3	4,195.9 ± 12.8	0.001057	0.279922 ± 06	1.46719 ± 2	1.88667 ± 3	−2.06 ± 0.26
S24b4	4,474.0 ± 0.8	4,470.1 ± 2.9	4,461.4 ± 9.9	0.001055	0.279906 ± 06	1.46719 ± 2	1.88667 ± 3	−1.57 ± 0.26
S24b7	4,473.9 ± 0.9	4,476.8 ± 2.8	4,483.3 ± 9.5	0.000742	0.279887 ± 05	1.46720 ± 1	1.88667 ± 3	−1.27 ± 0.19
S25b10	4,476.3 ± 0.9	4,474.5 ± 3.6	4,470.6 ± 12.5	0.001191	0.279926 ± 07	1.46721 ± 2	1.88669 ± 4	−1.21 ± 0.30
Average						**1.46720 ± 2**	**1.88667 ± 3**	
91500-1					0.282308 ± 06	1.46721 ± 1	1.88670 ± 3	
91500-2					0.282314 ± 11	1.46718 ± 2	1.88674 ± 4	
91500-3					0.282311 ± 06	1.46718 ± 2	1.88670 ± 4	
91500-4					0.282311 ± 05	1.46718 ± 1	1.88668 ± 2	
91500-5					0.282309 ± 05	1.46719 ± 1	1.88669 ± 2	
91500-6					0.282309 ± 06	1.46718 ± 1	1.88664 ± 2	
91500-7					0.282317 ± 05	1.46718 ± 1	1.88667 ± 3	
Average					**0.282311 ± 06**	**1.46718 ± 2**	**1.88669 ± 6**	

Age uncertainties are 2σ. Hf isotope ratios are reported normalized to the composition of the JMC-475 Hf standard. Uncertainties on the Hf isotope ratios reflect the 2-standard-error internal precision in the last decimal places. The external reproducibility of the $^{176}Hf/^{177}Hf$ ratio is estimated to be 22 p.p.m., based on the analyses of the seven 91500 zircon aliquots. U–Pb data are reported in full in Supplementary Table 1.

the basaltic breccia NWA 7034/7533 have been interpreted as igneous in origin[4,5]. Both the ages and the morphologies of the zircons analysed here are in line with this interpretation. Although the range of ages we report is consistent with earlier studies, the better than tenfold improvement in precision allows us to establish that zircon formation occurred in multiple igneous events over around 50 Myr. Four zircons analysed here define an age cluster at about 4,475 Myr (Fig. 1), which is considerably older than the age of about 4,430 Myr inferred from earlier studies of Martian zircons. In particular, one concordant zircon (S25B10) from this population records an age of 4,476.3 ± 0.9 Myr and, as such, represents the oldest directly dated material from Mars. This age is about 100 Myr older than the oldest dated terrestrial zircons[17], implying that the record of crust-forming processes on Mars is clearly older than that on Earth. These zircons thus provide a unique insight into the earliest history of the planet.

The distinct geochemical behaviour of Lu and Hf during partial melting episodes makes the ^{176}Lu–^{176}Hf decay system a powerful tool with which to constrain the timing of planetary silicate differentiation. For example, a primordial crustal reservoir will inherit a sub-chondritic $^{176}Lu/^{177}Hf$ ratio (<0.0336, ref. [18]) such that its time-integrated Hf isotopic composition will be less radiogenic than the chondritic uniform reservoir. We have determined the Hf isotope composition and Lu/Hf ratios of the seven U–Pb dated zircons using multiple collection inductively coupled plasma source mass spectrometry (MC-ICPMS)[19,20]. The more than fourfold improvement in precision afforded by this method over in situ techniques (see, for example, ref. [21]) is required to probe early differentiation timescales, given the small amount of radiogenic ingrowth of ^{176}Hf in the first 100 Myr of the Solar System. Moreover, this approach guarantees that the Hf isotope compositions were measured for the same volume of zircon for which the ages were determined, thereby ensuring that the initial Hf isotope compositions are accurately time-corrected. Importantly, the age variability of about 50 Myr recorded by the zircons allows us to track the isotopic evolution of their source reservoir. The seven zircons have unradiogenic initial εHf values ranging from −0.71 ± 0.32 to −2.06 ± 0.26 (Table 1; the εHf value is the deviation of the $^{176}Hf/^{177}Hf$ ratio of a sample from the chondritic uniform reservoir, in parts per 10^4), indicating that these grains formed from a precursor reservoir with a sub-chondritic $^{176}Lu/^{177}Hf$ ratio (Fig. 2a). The initial εHf values are correlated with their $^{207}Pb/^{206}Pb$ ages, suggesting that these grains were ultimately derived from a common source reservoir. The stable $^{178}Hf/^{177}Hf$ and $^{180}Hf/^{177}Hf$ ratios of the NWA 7034 zircons are identical to those of terrestrial zircons (Table 1), establishing that the reported εHf values are unaffected by neutron-capture effects.

The minimum model formation age for the source reservoir of the zircon population investigated here can be estimated from the oldest grain that records the most unradiogenic initial Hf isotope composition. Three of the four zircons that define the age cluster at about 4,475 Myr ago have concordant U–Pb ages and, as such, are inferred to have undisturbed ^{176}Lu–^{176}Hf systematics. Indeed, these three zircons (S24b4, S24b7 and S25b10) record identical initial εHf values within uncertainty despite having different $^{176}Lu/^{177}Hf$ ratios. Therefore, the average initial εHf value of these three concordant grains (−1.35 ± 0.22) provides a robust estimate of the Hf isotope composition of their protolith at around 4,475 Myr ago and, hence, the formation age of their source crustal reservoir. Given that the available geochemical data suggest that the bulk of the exposed Martian crust is of basaltic composition[13,22], we assess whether this type of crust could represent the source reservoir of the ancient zircon population. Using a $^{176}Lu/^{177}Hf$ ratio typical of mafic crustal sources on Earth (0.020, ref. [23]) returns impossibly old ages (>4,567 Myr), requiring the existence of a crustal reservoir with a composition more evolved than basaltic to account for the initial εHf value of the approximately 4,475-Myr-old zircons.

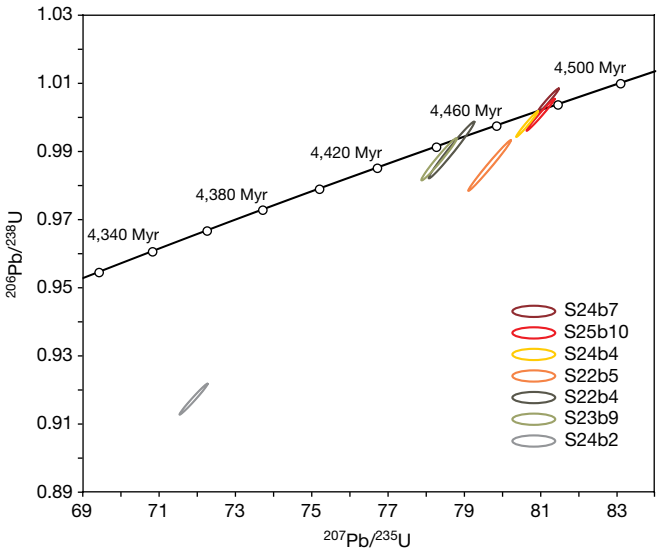


Fig. 1 | U–Pb concordia diagram for seven zircon grains from the NWA 7034 meteorite. Labels on the concordia curve represent time before present. Data point error ellipses are 2σ. Data used in this figure are reported in full in Supplementary Table 1.

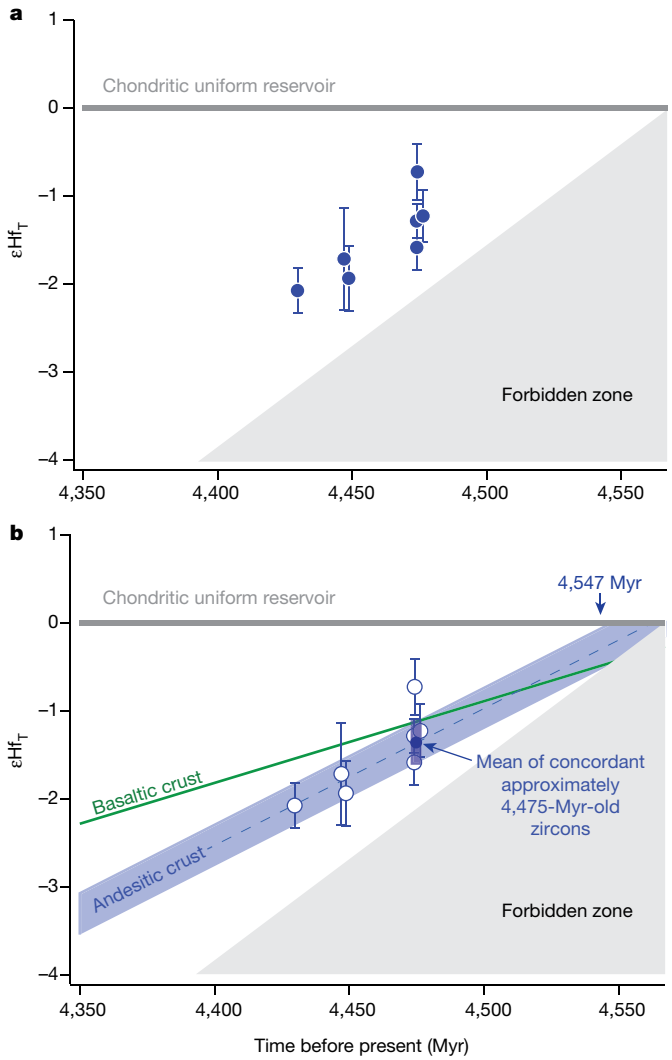


Fig. 2 | Hf isotope evolution diagrams. a, The initial εHf values (εHf$_T$) for the seven individual NWA 7034 zircons calculated with their corresponding ^{207}Pb–^{206}Pb ages using a λ^{176}Lu value of $1.867 \pm 0.008 \times 10^{-11}$ yr^{-1} (ref. [30]) and chondritic uniform reservoir parameters of ref. [18]. The upper boundary of the forbidden region represents a reservoir with ^{176}Lu/^{177}Hf = 0 and a formation age defined by the age of the Solar System[31] at 4,567 Myr such that no data should plot in this field. **b**, The time evolution of basaltic and andesitic crustal reservoirs required to account for the average initial Hf isotope compositions of the three concordant 4,475-Myr-old zircons (S24b4, S24b7 and S25b10) using ^{176}Lu/^{177}Hf ratios of 0.020 and 0.011 for the basaltic and andesitic crusts, respectively[23,25]. Considering the upper uncertainty of the zircon average εHf$_T$ value (-1.35 ± 0.22), it is not possible to account for the initial Hf isotope composition of these grains if they formed from the reworking of a basaltic crust, because extraction ages older than the Solar System are required. By contrast, using a more evolved, andesite-like ^{176}Lu/^{177}Hf ratio returns a minimum extraction age of 4,547 Myr. Using the mean of the concordant grains at face value and a ^{176}Lu/^{177}Hf ratio of 0.011 yields an extraction age of 4, 562^{+5}_{-15} Myr. We note that the time evolution of this reservoir can account for the Hf isotope composition of the younger, approximately 4,450-Myr-old and 4,430-Myr-old zircons. Indeed, a regression of the mean of the approximately 4,475-Myr-old, 4,450-Myr-old and 4,430-Myr-old zircons yields a slope corresponding to an andesite-like ^{176}Lu/^{177}Hf ratio of 0.011. Uncertainty on the εHf values reflects the internal precision (2 standard errors) or the external reproducibility of 22 p.p.m., whichever is larger. Uncertainty on the ^{207}Pb/^{206}Pb ages (2σ) are smaller than symbols.

One possibility is an andesite-like source, because rocks with such evolved compositions have been identified on Mars based on in situ observations[24]. Moreover, some magma ocean crystallization models[6,7] predict basaltic-to-andesitic compositions for the primary Martian crust produced by decompression melting of rising cumulates, following overturn of the gravitationally unstable stratified mantle. We note that recent estimates that suggest a low crustal bulk density for Mars ($2{,}582 \pm 209$ kg m^{-3}, ref. [25]) could, in principle, also be consistent with a more evolved average crustal composition. Using an andesite-like ^{176}Lu/^{177}Hf ratio of about 0.011 estimated from terrestrial rocks[23,26] defines a minimum formation age of 4,547 Myr for the source reservoir. An andesite-like composition for the nature of this source reservoir is further reinforced by the observation that the initial Hf isotope compositions of the approximately 4,450-Myr-old and 4,430-Myr-old zircons are also consistent with extraction from the same source (Fig. 2b). Indeed, taking the data at face value for each of the age groups returns a slope corresponding to an andesite-like ^{176}Lu/^{177}Hf ratio of about 0.011. The fact that more evolved compositions for a primordial crust are not predicted by any model provides confidence that the formation age of this reservoir cannot be younger than 4,547 Myr. Therefore, this minimum age for the source of the NWA 7034 zircons represents the oldest differentiated silicate reservoir yet identified on Mars. Ancient Martian zircons with ages comparable to that reported here have been identified in igneous, evolved lithologies that are interpreted to reflect re-melted primary Martian crust[4]. The enriched composition for the NWA 7034 zircon source reservoir inferred from the andesite-like ^{176}Lu/^{177}Hf ratio is consistent with this interpretation. Thus, our data require that a primordial crust existed on Mars by 4,547 Myr ago and that it survived for about 100 Myr before it was reworked to produce magmas, possibly by impacts[4,5], from which the NWA 7034 zircons crystallized. We infer that this primordial crust represents a global reservoir given its longevity and the extended period of reworking indicated by the zircon data.

The new timescales reported here for stabilization of the primordial Martian crust have far-reaching implications for understanding the accretion and differentiation history of Mars. Given that the formation of a stable primordial crust is the end product of the initial planetary differentiation, our data require that accretion, core formation and magma ocean crystallization on Mars was completed within 20 Myr of the formation of the Solar System. Such short timescales for primary accretion are predicted by planet formation models invoking pebble accretion where growth is fuelled by the gas-drag-assisted accretion of millimetre-sized objects, which leads to the efficient formation of Mars-sized embryos within the approximately 5-Myr lifetime of the protoplanetary disk[27,28]. Moreover, these timescales are also consistent with estimates based on the short-lived ^{182}Hf–^{182}W decay system for the timing of core formation, which is inferred to have occurred within 10 Myr of the formation of the Solar System[29]. By contrast, some recent studies have suggested that magma ocean crystallization was protracted on Mars, perhaps lasting up to 100 Myr after differentiation of the planet, on the basis of model ages deduced from the abundances of short-lived radionuclides in young Martian meteorites[10,15,16]. Such a protracted magma ocean crystallization is inconsistent with the data presented here and with thermal models suggesting that the solidification history of Mars must have been completed within about 10 Myr of accretion[7]. As such, the timing of silicate differentiation inferred from short-lived radionuclides in young Martian meteorites may not reflect primary differentiation of the planet but rather a younger, mantle-scale, fractionation event. Finally, our data and interpretation are fully consistent with models suggesting rapid magma ocean crystallization leading to a gravitationally unstable stratified mantle, which subsequently overturns, resulting in decompression melting of rising cumulates and extraction of a primordial basaltic-to-andesitic crust[6,7]. The extensive resurfacing of Mars by volcanism over the planet's history suggests that if any primordial crust is preserved, it will be deeply buried and may only be exposed in deep craters.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0222-z.

Received: 3 March 2018; Accepted: 1 May 2018;
Published online 27 June 2018.

1. Stevenson, R. K. & Patchett, P. J. Implications for the evolution of continental-crust from Hf-isotope systematics of Archean detrital zircons. *Geochim. Cosmochim. Acta* **54**, 1683–1697 (1990).
2. Amelin, Y., Lee, D.-C., Halliday, A. N. & Pidgeon, R. T. Nature of the Earth's earliest crust from hafnium isotopes in single detrital zircons. *Nature* **399**, 252–255 (1999).
3. Amelin, Y., Lee, D.-C. & Halliday, A. N. Early-middle Archaean crustal evolution deduced from Lu-Hf and U-Pb isotopic studies of single grain zircons. *Geochim. Cosmochim. Acta* **64**, 4205–4225 (2000).
4. Humayun, M. et al. Origin and age of the earliest Martian crust from meteorite NWA 7533. *Nature* **503**, 513–516 (2013).
5. McCubbin, F. M. et al. Geologic history of Martian regolith breccia Northwest Africa 7034: evidence for hydrothermal activity and lithologic diversity in the Martian crust. *J. Geophys. Res. Planets* **121**, 2120–2149 (2016).
6. Elkins-Tanton, L. T., Hess, P. C. & Parmentier, E. M. Possible formation of ancient crust on Mars through magma ocean processes. *J. Geophys. Res. Planets* **110**, E12S01 (2005).
7. Elkins-Tanton, L. T. Linked magma ocean solidification and atmospheric growth for Earth and Mars. *Earth Planet. Sci. Lett.* **271**, 181–191 (2008).
8. Dauphas, N. & Pourmand, A. Hf–W–Th evidence for rapid growth of Mars and its status as a planetary embryo. *Nature* **473**, 489–492 (2011).
9. Nimmo, F. & Tanaka, K. Early crustal evolution of Mars. *Annu. Rev. Earth Planet. Sci.* **33**, 133–161 (2005).
10. Borg, L. E., Brennecka, G. A. & Symes, S. J. K. Accretion timescale and impact history of Mars deduced from the isotopic systematics of martian meteorites. *Geochim. Cosmochim. Acta* **175**, 150–167 (2016).
11. Caro, G. Early silicate Earth differentiation. *Annu. Rev. Earth Planet. Sci.* **39**, 31–58 (2011).
12. Carlson, R. W. et al. How did early Earth become our modern world? *Annu. Rev. Earth Planet. Sci.* **42**, 151–178 (2014).
13. McSween, H. Y. Petrology on Mars. *Am. Mineral.* **100**, 2380–2395 (2015).
14. Schiller, M., Bizzarro, M. & Fernandes, V. A. Isotopic evolution of the protoplanetary disk and the building blocks of Earth and Moon. *Nature* **555**, 507–510 (2018).
15. Debaille, V., Brandon, A. D., Yin, Q.-Z. & Jacobsen, B. Coupled ^{142}Nd–^{143}Nd evidence for a protracted magma ocean in Mars. *Nature* **450**, 525–528 (2007).
16. Kruijer, T. S. et al. The early differentiation of Mars inferred from Hf-W chronometry. *Earth Planet. Sci. Lett.* **474**, 345–354 (2017).
17. Whitehouse, M. J., Nemchin, A. A. & Pidgeon, R. T. What can Hadean detrital zircon really tell us? A critical evaluation of their geochronology with implications for the interpretation of oxygen and hafnium isotopes. *Gondwana Res.* **51**, 78–91 (2017).
18. Bouvier, A., Vervoort, J. D. & Patchett, P. J. The Lu–Hf and Sm–Nd isotopic composition of CHUR: constraints from unequilibrated chondrites and implications for the bulk composition of terrestrial planets. *Earth Planet. Sci. Lett.* **273**, 48–57 (2008).
19. Bizzarro, M., Baker, J. A. & Ulfbeck, D. A new digestion and chemical separation technique for rapid and highly reproducible determination of Lu/Hf and Hf isotope ratios in geological materials by MC-ICP-MS. *Geostand. Newsl.* **27**, 133–145 (2003).
20. Connelly, J. N., Ulfbeck, D. G., Thrane, K., Bizzarro, M. & Housh, T. A method for purifying Lu and Hf for analysis by MC-ICP-MS using TODGA resin. *Chem. Geol.* **233**, 126–136 (2006).
21. Kemp, A. I. S. et al. Hadean crustal evolution revisited: new constraints from Pb-Hf isotope systematics of the Jack Hills zircons. *Earth Planet. Sci. Lett.* **296**, 45–56 (2010).
22. McSween, H. Y., Taylor, J. & Wyatt, M. B. Elemental composition of the Martian crust. *Science* **324**, 736–739 (2009).
23. Rudnick, R. L. & Gao, S. in *The Crust* (ed. Rudnick, R. L.) *Treatise on Geochemistry* Vol. 3, 1–64 (Elsevier, Amsterdam, 2003).
24. Sautter, V. et al. In situ evidence for continental crust on early Mars. *Nat. Geosci.* **8**, 605–609 (2015).
25. Goossens, S. et al. Evidence for a low bulk crustal density for Mars from gravity and topography. *Geophys. Res. Lett.* **44**, 7686–7694 (2017).
26. Condie, K. C. Chemical composition and evolution of the upper continental crust: contrasting results from surface samples and shales. *Chem. Geol.* **104**, 1–37 (1993).
27. Johansen, A., Mac Low, M. M., Lacerda, P. & Bizzarro, M. Growth of asteroids, planetary embryos, and Kuiper belt objects by chondrule accretion. *Sci. Adv.* **1**, e1500109 (2015).
28. Bollard, J. et al. Early formation of planetary building blocks inferred from Pb isotopic ages of chondrules. *Sci. Adv.* **3**, e1700407 (2017).
29. Mezger, K., Debaille, V. & Kleine, T. Core formation and mantle differentiation on Mars. *Space Sci. Rev.* **174**, 27–48 (2013).
30. Söderlund, U., Patchett, P. J., Vervoort, J. D. & Isachsen, C. E. The ^{176}Lu decay constant determined by Lu-Hf and U-Pb isotope systematics of Precambrian mafic intrusions. *Earth Planet. Sci. Lett.* **219**, 311–324 (2004).
31. Connelly, J. N. et al. The absolute chronology and thermal processing of solids in the solar protoplanetary disk. *Science* **338**, 651–655 (2012).

Acknowledgements Financial support for this project was provided by the Danish National Research Foundation (DNRF97) and the European Research Council (ERC Consolidator Grant Agreement 616027, STARDUST2ASTEROIDS) to M.B. We thank J. Frydenvang and K. Kinch for discussions.

Reviewer information *Nature* thanks A. Brandon and L. Elkins-Tanton for their contribution to the peer review of this work.

Author contributions M.B. designed and led the research project. M.M.C. and J.N.C. identified and separated the zircons and performed analytical work related to the U–Pb isotope systematics of the zircons. L.C.B., J.N.C. and M.B. performed analytical work related to the ^{176}Lu–^{176}Hf systematics of the zircons. N.K.J., D.W., M.S., M.J.W., J.F.S., J.J.B., A.A.N., F.M., A.A. and B.G. assisted in sample preparation and zircon identification. All authors participated in the interpretation of the data. The manuscript was written by L.C.B., M.M.C., J.N.C. and M.B.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0222-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0222-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

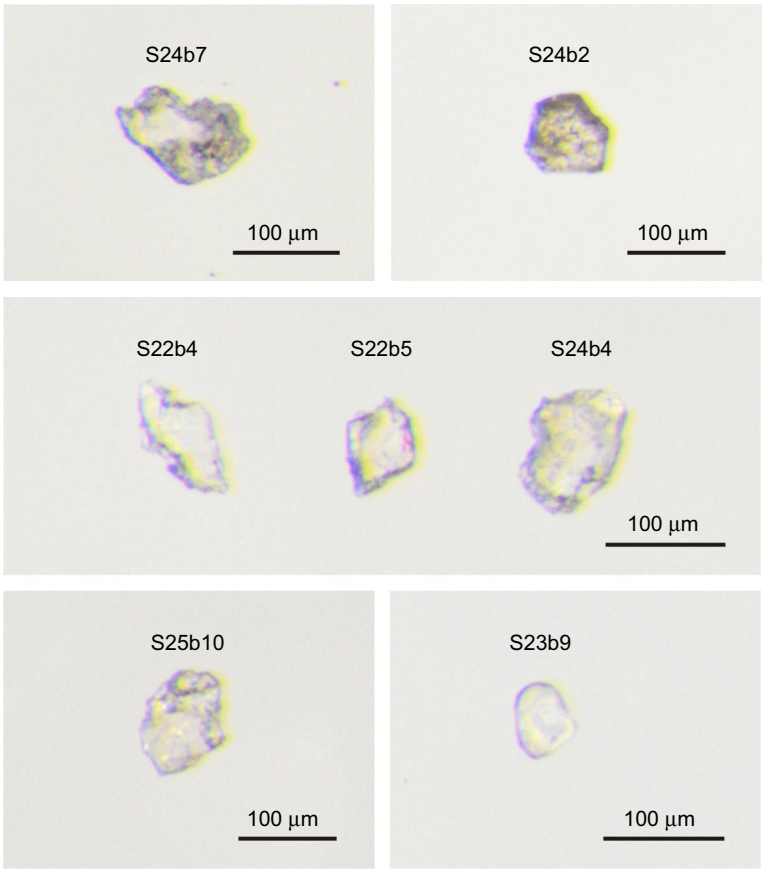
A total of seven zircon grains were extracted from a crushed bulk rock aliquot of NWA 7034 and analysed for U–Pb and Lu–Hf systematics. Given the limited number of zircons recovered from the crushing process and their small sizes, only one of the larger grains (S25b10) was chemically abraded[32]. This pre-treatment consisted of thermally annealing the less metamict domains of the crystal over three days at 900 °C, followed by dissolution of the un-annealed portions using concentrated HF for 12 h at 180 °C in Teflon capsules. Before complete dissolution, all zircon grains were cleaned in Pyrex beakers in an ultrasonic bath with alternating steps of warm 3.5 M HNO_3, H_2O and acetone. As we assumed that each grain might represent a different age, the grains were processed as single grains. The individual crystals were dissolved in separate PFE Teflon capsules in a HF:HNO_3 (3:1) mixture, together with the mixed ^{202}Pb–^{205}Pb–^{233}U–^{235}U EARTHTIME U–Pb tracer[33], for four days at 210 °C. The dissolved samples were evaporated to dryness and redissolved in 3.1 M HCl overnight. Uranium and lead were separated from the matrix elements by anion chromatography using 50-μl Teflon columns[34,35] and dried down together with 8 μl of 0.1 M H_3PO_4. They were loaded with silica gel[36] on previously outgassed zone-refined Re filaments. The Pb and U isotopic ratios of the sample + tracer mixture were measured using the Triton Thermo-Fisher thermal ionization mass spectrometer at the Centre for Star and Planet Formation, University of Copenhagen, where each isotope was sequentially counted in a single axial ion counting system with Pb as Pb^+, and U as UO^{2+}. The data were reduced offline and instrumental mass fractionation was accounted for by a linear mass-dependent fractionation law based on the $^{202}Pb/^{205}Pb$ ratio of the tracer. After 1 pg of laboratory Pb blank was removed from the analyses, the remainder of common Pb was assumed to have an isotopic composition, modelled after ref. [37]. Instrumental mass-dependent fractionation of U was accounted for using the $^{233}U/^{235}U$ ratio of the tracer, which included a correction for the isobaric interference of $^{233}U^{16}O^{18}O$ on the $^{235}U^{16}O_2$ peak at mass 267. The $^{238}U/^{235}U$ ratio of mass 137.88 (ref. [38]) and the ^{238}U and ^{235}U decay constants of ref. [39] were used for calculation of U/Pb ratios and ages. All ratios and age uncertainties are quoted at the 95% confidence level.

The Hf isotope composition and Lu/Hf ratios of individual zircons were determined from the same sample digestion as that used for U–Pb age determination. Following collection of the high-field-strength element and rare-earth element (REE) washes from the U–Pb purification, approximately 5% of the solution was aliquoted for Lu/Hf ratio determination. The Hf was purified from the remaining solution by a two-step procedure using TEVA-spec and TODGA resins (Eichrom Industries) based on protocols outlined in refs 19 and 20. Zr was quantitatively separated from Hf, Ti and REE using 100–150 μm TEVA-spec resin in a 120-μl column. The fractions containing high-field-strength elements and REE were loaded in 0.6 ml of 10.5 M HCl and a Hf + Ti + REE cut was successively collected with 3.6 ml of 10.5 M HCl and 4.2 ml of 9.5 M HCl. Zr was recovered with 3.0 ml of 6 M HCl. Hf was purified from Ti and REE using 50–100 μm TODGA resin in 200-μl columns. The Hf + Ti + REE fractions were loaded in 0.85 ml of 3.5 M HNO_3-0.06 M boric acid and the Ti was eluted by adding 3.5 ml of 3.5 M HNO_3 while Hf and REE remained on the column. Hf was subsequently collected with 6 ml of 1 M HNO_3-0.35 M HF. This method returns Hf yields greater than 95% in Hf cuts with Zr/Hf ratios below 1. Hf isotope ratios were measured on the Pandora Thermo-Fisher Neptune Plus MC-ICPMS at the Centre for Star and Planet Formation, University of Copenhagen, using a sample-standard bracketing technique. Samples were aspirated into the plasma source in 2% HNO_3-0.1 M HF solution via a Cetac Aridus II desolvating nebulizer using Ar and N_2 sweep gases with an uptake rate of about 100 μl min^{-1}. Typical sensitivity at this uptake rate was 1,800 V per p.p.m. for Hf. Hafnium isotopic data were acquired in static mode using eight Faraday collectors allowing for simultaneous measurement of ^{176}Hf, ^{177}Hf, ^{178}Hf, ^{179}Hf and ^{180}Hf as well as monitoring potential isobaric interferences (^{176}Yb on ^{176}Hf, ^{176}Lu on ^{176}Hf and ^{180}W on ^{180}Hf) using ^{175}Lu, ^{171}Yb and ^{182}W. Faraday detectors used to collect Hf isotopes and ^{182}W were connected to amplifiers with 10^{11}-Ω feedback resistors, whereas ^{175}Lu and ^{171}Yb were measured on Faraday detectors connected to amplifiers with 10^{13}-Ω feedback resistors. Sample analyses were interspersed with analyses of the JMC-475 standard as follows: JMC-475 (1), JMC-475 (2), sample-1, JMC-475 (3), JMC-475 (4). Samples and standards were analysed with a signal intensity of at least 1 V on mass ^{177}Hf and ensuring that the signal intensity of the sample and standard were matched to within 5%. Samples were analysed once and the total amount of Hf consumed per analysis was typically 2–5 ng for NWA 7034 zircons. Total procedural blanks were <10 pg for Hf, an amount that is negligible for all samples considering the amount of Hf available for analysis. All data reduction was conducted off-line using the freely available Iolite data reduction software[40] that runs within Igor Pro. Background intensities were interpolated using a smoothed cubic spline, as were changes in mass bias with time. Hf isotope data were corrected for mass bias using the exponential mass fractionation law, adopting $^{179}Hf/^{177}Hf = 0.7325$. The sample $^{176}Hf/^{177}Hf$, $^{178}Hf/^{177}Hf$ and $^{180}Hf/^{177}Hf$ ratios were normalized to the JMC-475 reference values of 0.282160, 1.46717 and 1.88667, respectively. Contributions from interfering species were on average 33 p.p.m. and 1 p.p.m. on $^{176}Hf/^{177}Hf$ from Yb and Lu, respectively, and 26 p.p.m. on $^{180}Hf/^{177}Hf$ from W. Doping experiments with Yb interference levels at least ten times greater than typically observed in our sample demonstrate that our interference correction is accurate. The accuracy and external reproducibility of our method was assessed by repeat analyses of the 91500 zircon reference standard[41]. In detail, seven aliquots of the same sample dissolution each containing approximately 5 ng of Hf were individually processed through our U–Pb and Hf purification scheme and analysed following the methods described above. The average values obtained for the $^{176}Hf/^{177}Hf$, $^{178}Hf/^{177}Hf$ and $^{180}Hf/^{177}Hf$ ratios of the 91500 standard aliquots were 0.282311 ± 0.000006, 1.46718 ± 0.00002 and 1.88669 ± 0.00006, where the uncertainty represents the external reproducibility (2 standard deviations) (Table 1). The data we obtained for the 91500 zircon reference standard are identical to published values[42].

The Lu/Hf ratios were determined using the same MC-ICPMS as used for the Hf isotopic measurements using gravimetrically prepared mixed Lu–Hf standard solutions. In detail, the approximately 5% aliquot reserved for Lu/Hf ratio determination was evaporated and re-dissolved in 2% HNO_3-0.1 M HF prior to analysis. Samples were aspirated into the plasma source via a Cetac Aridus II desolvating nebulizer using Ar and N_2 sweep gases with an uptake rate of about 100 μl min^{-1}. The ^{175}Lu beam was collected on the axial secondary electron multiplier, whereas the ^{177}Hf was collected on the H2 Faraday detector connected to an amplifier with a 10^{11}-Ω feedback resistor. Sample analyses were interspaced by analyses of the calibrated Lu–Hf standard solution as follows: Lu–Hf standard (1), Lu–Hf standard (2), sample-1, Lu–Hf standard (3), Lu–Hf standard (4). Samples and standard were analysed with an intensity of at least 0.05 V on mass ^{177}Hf and 100.000 counts per second on mass ^{175}Lu, ensuring that the signal intensity of the sample and standard were matched to within about 5%. Total procedural blanks were <5 fg for Lu and negligible for all samples considering the amount of Lu available for analysis. The Lu–Hf standard solution was prepared gravimetrically to match the typical Lu/Hf ratio of zircon and is accurate to 2%. The external reproducibility of our approach was estimated by repeated analysis of the 91500 zircon standard. Analysis of ten individual aliquots of a single dissolution of the 91500 zircon standard yielded a $^{176}Lu/^{177}Hf$ ratio of 0.00030346 ± 0.0000016 (2 standard deviations), which corresponds to an external reproducibility of 0.5% for the Lu/Hf ratio. Potential fractionation of the Lu/Hf ratio induced by U–Pb purification was evaluated by measuring the Lu/Hf ratios of aliquots of the 91500 zircon standard before and after U–Pb purification. Our tests demonstrate that potential fractionation of the Lu/Hf during U–Pb purification is less than 0.4%. Combining this with the uncertainty of the Lu–Hf standard solution, the external reproducibility of 0.5% and the potential fractionation of 0.4%, we infer an accuracy of 2.1% for our Lu/Hf ratio measurements. This represents the total uncertainty on the Lu/Hf ratio reported here and has been propagated in the final uncertainties quoted for the initial Hf isotope composition of the NWA 7034 zircons.

Data availability. The authors declare that data supporting the findings of this study are available within the paper and the Methods. All other data are available from the corresponding author upon reasonable request.

32. Mattinson, J. M. et al. Zircon U-Pb chemical abrasion ("CA-TIMS") method: combined annealing and multi-step partial dissolution analysis for improved precision and accuracy of zircon ages. *Chem. Geol.* **220**, 47–66 (2005).
33. Condon, D. J., Schoene, B., McLean, N. M., Bowring, S. A. & Parrish, R. R. Metrology and traceability of U–Pb isotope dilution geochronology (EARTHTIME Tracer Calibration Part I). *Geochim. Cosmochim. Acta* **164**, 464–480 (2015).
34. Krogh, T. E. A low contamination method for hydrothermal decomposition of zircon and extraction of U and Pb for isotopic age determination. *Geochim. Cosmochim. Acta* **37**, 485–494 (1973).
35. Corfu, F. U–Pb age, setting and tectonic significance of the anorthosite–mangerite–charnockite–granite suite, Lofoten-Vesterålen, Norway. *J. Petrol.* **56**, 2081–2097 (2004).
36. Gerstenberger, H. & Haase, G. A highly effective emitter substance for mass spectrometric Pb isotope ratio determinations. *Chem. Geol.* **136**, 309–312 (1997).
37. Bellucci, J. J. et al. Pb-isotopic evidence for an early, enriched crust on Mars. *Earth Planet. Sci. Lett.* **410**, 34–41 (2015).
38. Steiger, R. H. & Jager, E. Subcommission on geochronology: convention on the use of decay constants in geo- and cosmochronology. *Earth Planet. Sci. Lett.* **36**, 359–362 (1977).
39. Jaffey, A. H., Flynn, K. F., Glendenin, L. E., Bentley, W. C. & Essling, A. M. Precision measurement of half-lives and specific of ^{235}U and ^{238}U. *Phys. Rev. C* **4**, 1889–1906 (1971).
40. Paton, C., Hellstrom, J., Paul, B., Woodhead, J. & Hergt, J. Iolite: freeware for the visualisation and processing of mass spectrometric data. *J. Anal. At. Spectrom.* **26**, 2508–2518 (2011).
41. Wiedenbeck, M. et al. Three natural zircon standards for U-Th-Pb, Lu-Hf, trace element and REE analyses. *Geostand. Newsl.* **19**, 1–23 (1995).
42. Blichert-Toft, J. Hf isotopic composition of zircon reference material 91500. *Chem. Geol.* **253**, 252–257 (2008).



Extended Data Fig. 1 | Photomicrographs of the NWA 7034 zircons analysed in this study taken under natural light. Given the small size and limited number of zircons recovered from the crushing process, we considered it to be preferable not to conduct additional imaging (using cathodoluminescence) because this necessitates extra manipulation of the individual grains, thereby increasing the risk of losing zircons. The fact that the zircons have mostly concordant U–Pb ages confirms their simple igneous history and, therefore, additional imaging to investigate potential zoning is not required here.

LETTER

https://doi.org/10.1038/s41586-018-0244-6

A synaptic threshold mechanism for computing escape decisions

Dominic A. Evans[1,2,3], A. Vanessa Stempel[1,2,3], Ruben Vale[1,2], Sabine Ruehle[1,2], Yaara Lefler[1,2] & Tiago Branco[1,2]*

Escaping from imminent danger is an instinctive behaviour that is fundamental for survival, and requires the classification of sensory stimuli as harmless or threatening. The absence of threat enables animals to forage for essential resources, but as the level of threat and potential for harm increases, they have to decide whether or not to seek safety[1]. Despite previous work on instinctive defensive behaviours in rodents[2–11], little is known about how the brain computes the threat level for initiating escape. Here we show that the probability and vigour of escape in mice scale with the saliency of innate threats, and are well described by a model that computes the distance between the threat level and an escape threshold. Calcium imaging and optogenetics in the midbrain of freely behaving mice show that the activity of excitatory neurons in the deep layers of the medial superior colliculus (mSC) represents the saliency of the threat stimulus and is predictive of escape, whereas glutamatergic neurons of the dorsal periaqueductal grey (dPAG) encode exclusively the choice to escape and control escape vigour. We demonstrate a feed-forward monosynaptic excitatory connection from mSC to dPAG neurons, which is weak and unreliable—yet required for escape behaviour—and provides a synaptic threshold for dPAG activation and the initiation of escape. This threshold can be overcome by high mSC network activity because of short-term synaptic facilitation and recurrent excitation within the mSC, which amplifies and sustains synaptic drive to the dPAG. Therefore, dPAG glutamatergic neurons compute escape decisions and escape vigour using a synaptic mechanism to threshold threat information received from the mSC, and provide a biophysical model of how the brain performs a critical behavioural computation.

Detecting and escaping from threats is an instinctive behaviour that reduces the chances of being harmed, but also results in the halting of other behaviours and the potential loss of resources. To balance escape with other survival behaviours, animals use sensory information and past experience to estimate threat and decide whether or not to escape[1]. Although perceptual decision-making has been studied in primates and rodents using learned-choice tasks[12,13], and previous work has identified key circuits for innate defence[4–8,14,15], the neurophysiological basis of escape decisions in mammals is largely unknown. Here we investigated escape in mice using innately aversive overhead expanding spots[3,16], while varying the spot contrast to manipulate the saliency of the stimulus. Presentation of the stimulus while mice explored an arena with a shelter resulted in shelter-directed escape responses that were variable and probabilistic (Fig. 1a–c). Decreasing the stimulus contrast progressively increased reaction times and reduced escape probability, producing chronometric and psychometric curves similar to those from learned perceptual categorisation tasks[12,13] (Fig. 1d, e, Supplementary Video 1). Response vigour (measured as the escape speed) also increased with contrast (Fig. 1f), showing that probability, reaction time and vigour of instinctive escape are innately matched to the saliency of the threat stimulus (see also Extended Data Fig. 1). The relationship between these variables was well described by a drift-diffusion model[12,17] that integrates a noisy threat level variable over time and implements the escape decision as a threshold-crossing process (Fig. 1g, see Methods). This was further supported by exposing mice to innately aversive ultrasonic sweeps, which generated escape with high probability, short reaction times and high vigour (Fig. 1b–f).

Multiple brain regions contribute to instinctive defensive behaviours[5,7,8,14,18,19], so we next used optogenetic inactivation[20] of excitatory neurons expressing vesicular glutamate transporter 2 (VGluT2$^+$) to define critical circuit nodes for escape (Fig. 2a, b). Inactivation of the dPAG and mSC both severely affected escape—without affecting exploratory behaviour (Extended Data Fig. 2)—but in different ways. The inactivation of dPAG neurons switched the response to threat

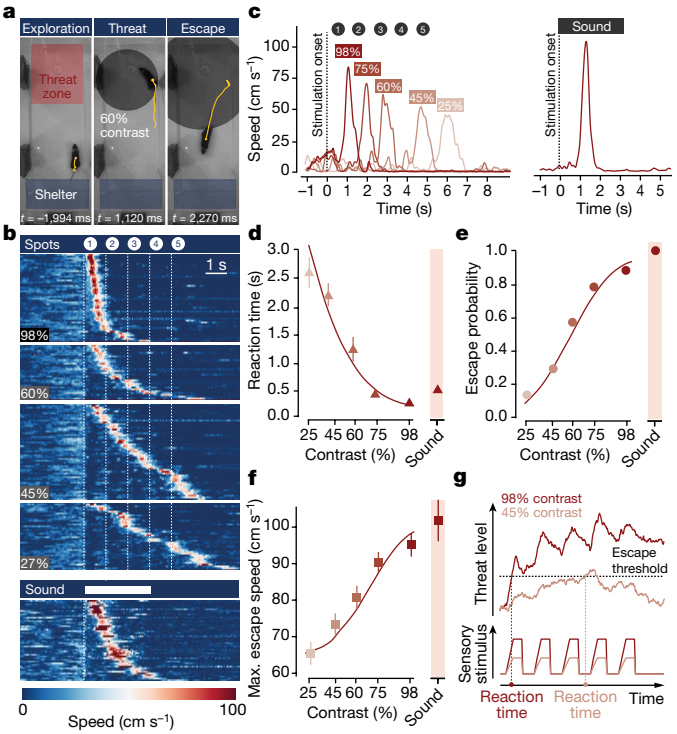


Fig. 1 | Escape behaviour during threats of varying intensity. a, Video frames of escape to expanding spots. Yellow lines show the trajectory of the mouse during the preceding 2 s, stimulation onset is $t = 0$. **b**, Raster plot of mouse speed during escape trials for visual (top, organized by contrast) and sound (bottom) stimulation, sorted by reaction time ($n = 13$ mice). **c**, Single trial traces from one mouse escaping from different contrast spots (left) and sound (right). **d**–**f**, Chronometric (**d**) psychometric (**e**) and vigour (**f**) curves of contrast and escape behaviour; $n = 13$ mice, 209 trials; escape probability: $P = 2.5 \times 10^{-7}$, reaction time: $P = 3.5 \times 10^{-8}$, vigour: $P = 1.6 \times 10^{-6}$. **g**, Theoretical model for computing escape from threat stimuli. Data points in **d**–**f** are means of trials pooled across mice, error bars are s.e.m., red lines are model fits to the data, P values are calculated using repeated-measures analysis of variance (ANOVA).

[1]MRC Laboratory of Molecular Biology, Cambridge, UK. [2]UCL Sainsbury Wellcome Centre for Neural Circuits and Behaviour, London, UK. [3]These authors contributed equally: D. A. Evans, A. V. Stempel. *e-mail: t.branco@ucl.ac.uk

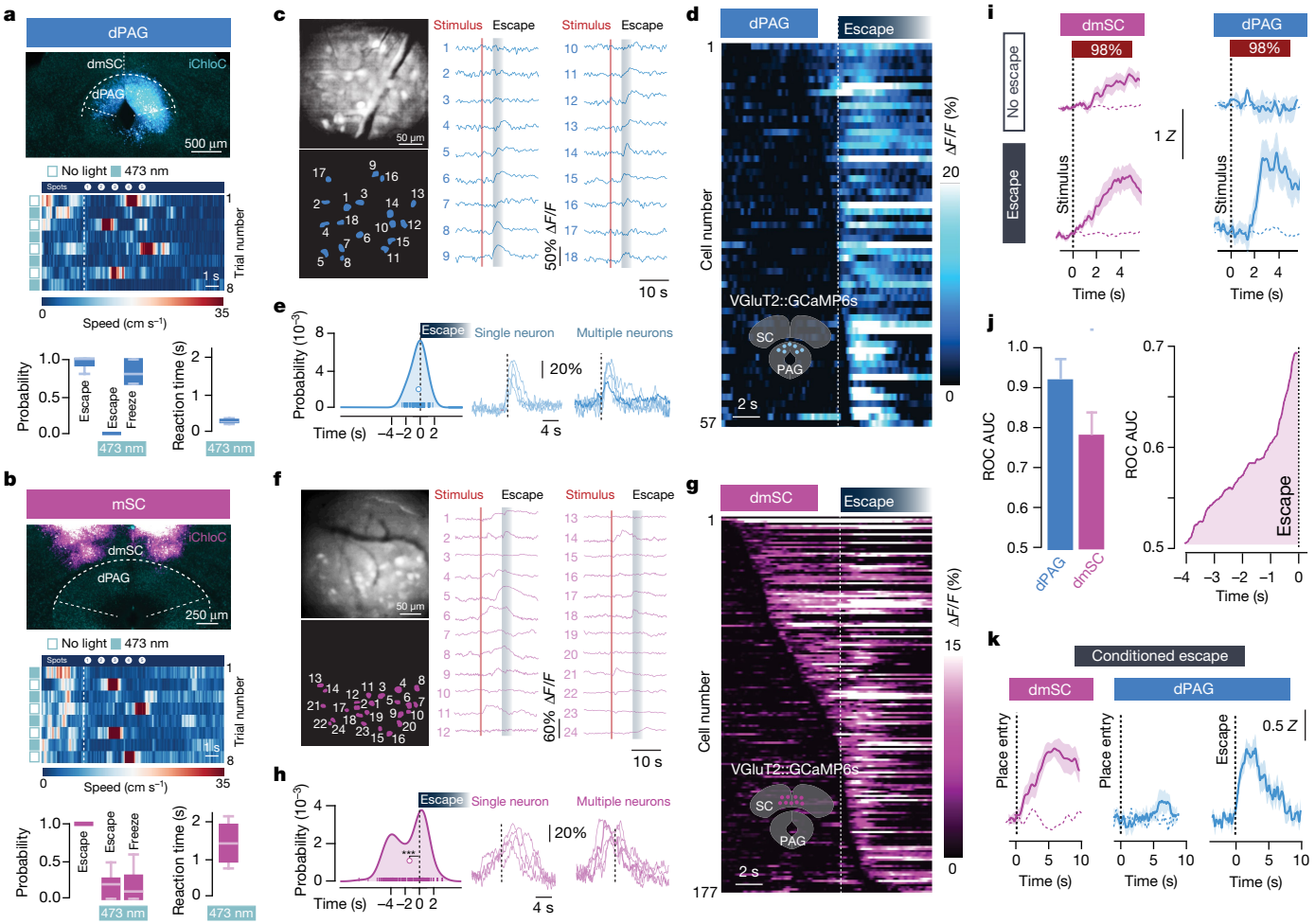


Fig. 2 | Encoding of threat and escape behaviour in the superior colliculus and periaqueductal gry. a, iChloC expression in VGluT2$^+$ dPAG neurons (top), speed raster during interleaved trials of threat presentation with light-off or on (middle), and summary for stimulation during dPAG inactivation (bottom) ($P_{escape} = 0.03 \pm 0.03$, $P_{freeze} = 0.86 \pm 0.06$, mean freezing duration $= 4.3 \pm 1.0$ s; $n = 6$ mice; escape: $P = 8.12 \times 10^{-5}$, freezing: $P = 0.00029$; *U*-tests between light-off and light-on). **b**, Same as **a** for VGluT2$^+$ mSC inactivation ($P_{escape} = 0.18 \pm 0.05$, $P_{freeze} = 0.19 \pm 0.07$; $n = 9$ mice; escape: $P = 5.15 \times 10^{-5}$, freezing: $P = 0.02$; *U*-tests as above). Reaction times are slower during mSC inactivation than during dPAG inactivation ($P = 0.002$, two-tailed *t*-test). **c**, Field-of-view of dPAG VGluT2::GCaMP6s neurons (top left), cell mask (bottom left) and single-trial examples (right). **d**, Average calcium response for active dPAG cells, aligned to escape and sorted by onset (57 out of 138 cells, $n = 3$ mice, 55 trials). **e**, Left, distribution of dPAG cell onsets (curve is kernel density estimation, markers show onsets). Mean onset $= -0.24 \pm 0.21$ s (white marker, not different from 0 s; $P = 0.24$, one-sample *t*-test). Right, example single-trial traces. **f**–**h**, Same as **c**–**e** for dmSC (177 out of 218 active cells, $n = 8$ mice, 111 trials; mean onset $= -1.51 \pm 0.17$ s, $P = 3.5 \times 10^{-12}$, Wilcoxon signed-rank test comparison with 0 s). **i**, Population activity for 98% contrast (*z*-score), grouped by trial outcome for dmSC (pink; 111 trials, $P = 0.023$, two-tailed *t*-test between escape and no escape; $P = 5.8 \times 10^{-10}$, one-sample *t*-test between no escape and 0 s) and dPAG (blue; 55 trials, $P = 0.00028$ and $P = 0.11$, tests as for mSC). Dashed lines are activity without stimulus. **j**, ROC area under the curve (AUC, left), and AUC evolution for dmSC signals up to escape (right, 75 trials; error bars are s.d.). **k**, dmSC activity upon place entry increases after conditioning (left, dashed line is activity before conditioning; 57 trials, $n = 7$ mice, $P = 0.00013$, two-tailed *t*-test between pre- and post-conditioning), whereas dPAG activity increases selectively upon escape (middle and right, *z*-score $= 1.5 \pm 0.2$, 20 trials, $n = 3$ mice, $P = 0.0004$, two-tailed *t*-test between pre- and post-conditioning). Box-and-whisker plots show median, interquartile range (IQR) and range. Error bars and shaded areas are s.e.m.; *** $P < 0.001$.

from escape to freezing, with fast reaction times (269 ± 35 ms, Fig. 2a; Supplementary Video 2), indicating that the threat was still detected and that the dPAG is specifically required to initiate escape. By contrast, visual and sound stimulation after mSC inactivation produced no defensive response in $62 \pm 10\%$ of light-on trials, which suggests that the link between the sensory stimulus and the response to threat was severely compromised (Fig. 2b, Supplementary Video 3). In the remaining trials, the reaction time was slow ($1{,}443 \pm 255$ ms, Fig. 2b) and the vigour of escape was reduced (Extended Data Fig. 2c), which is compatible with a reduction in the perceived level of threat. Similar results were obtained upon muscimol inactivation of the dPAG and mSC, whereas inactivation of the visual cortex (V1) or the amygdala caused only small decreases in escape probability and vigour (Extended Data Fig. 3). Next we performed calcium imaging of VGluT2$^+$ neurons in the deep layers of the mSC (dmSC) or in the dPAG in freely behaving mice. Activity in both areas increased during stimulus-evoked escape (Fig. 2c, f), with a trial reliability of $28 \pm 3\%$ for the dPAG and $35 \pm 3\%$ for the dmSC; this yielded a mean fraction of active cells of $14 \pm 5\%$ and $23 \pm 6\%$, respectively, which was stable over multiple trials (Extended Data Fig. 4). However, the temporal profile of dPAG and dmSC activity was distinct. Whereas dPAG cells were active in the peri-escape initiation period (Fig. 2d, e), activity in most dmSC cells preceded escape onset (Fig. 2g, h), and this temporal difference was also reflected in the ensemble activity onset (onset relative to the start of escape: -0.25 ± 0.48 s for dPAG, -1.77 ± 0.5 s for dmSC, $P = 0.59$ and $P = 0.00075$ respectively, two-tailed *t*-test comparison with escape onset). Sorting trials from the same stimulus contrast by trial outcome (Fig. 2i) showed that dmSC neurons encode the threat stimulus, and also reflect the choice to escape (*z*-score $= 1.93 \pm 0.23$ for escape, 1.18 ± 0.11 for no escape), whereas activity in dPAG neurons increases

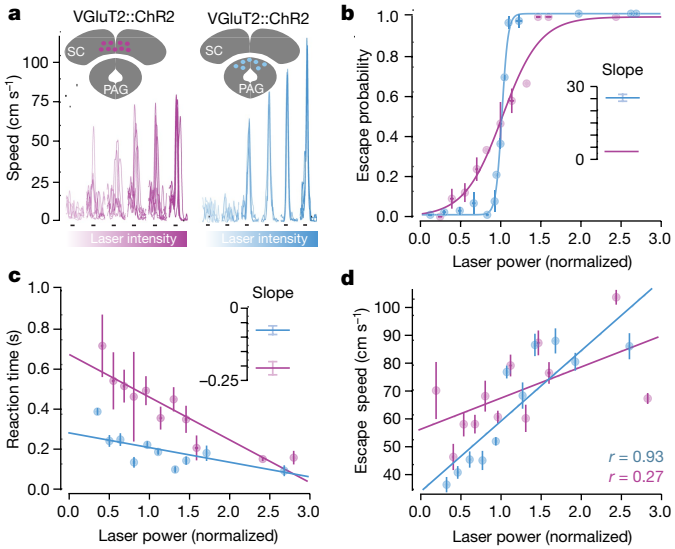


Fig. 3 | Optogenetic stimulation shows different roles for mSC and dPAG in escape behaviour. a, Speed traces with increasing light intensity (10 Hz pulse, black lines) from one mouse (left, mSC; right, dPAG). **b**, Psychometric curve (mSC: 278 trials, $n = 4$ mice, slope = 4.0, 95% confidence interval (2.75, 5.25)); dPAG: 590 trials, $n = 7$ mice, slope = 26.3, 95% confidence interval (22.1, 30.4)). Lines are logistic fits (pooled across all mice and binned light intensities), inset shows fit slope (error bars are s.d.). **c**, Chronometric curve (mSC: 149 trials, slope = −0.21, 95% confidence interval (−0.27, −0.15); dPAG: 328 trials, slope = −0.07, 95% confidence interval (−0.11, −0.03)). Lines are linear fits, inset as **b**. **d**, Correlation between light intensity and escape speed (mSC: 149 trials, $P = 0.04$; dPAG: 328 trials, $P = 1.5 \times 10^{-5}$; Pearson's r). Error bars are s.e.m. unless otherwise indicated, mSC data are shown in purple and dPAG in blue.

exclusively in escape trials (z-score = 2.28 ± 0.17 for escape, 0.49 ± 0.19 for no escape). Receiver–operator characteristic (ROC) analysis of ensemble activity reflected this difference, and showed that an ideal observer of dmSC activity could predict the decision to escape 900 ms before escape initiation (68% correct; Fig. 2j). Ensemble dmSC activity also showed a strong negative correlation with reaction time, further suggesting that it is important for escape initiation (Extended Data Fig. 4i, j). To test further the nature of dmSC signals, we exposed mice to a place-aversion paradigm that resulted in spontaneous flight upon approaching the threat area (Extended Data Fig. 5, Supplementary Video 4). The activity of dmSC neurons after conditioning increased upon place entry and preceding escape, despite there being no stimulus presentation (z-score = 1.94 ± 0.17; Fig. 2k). Importantly, pre-escape activity was still predictive of escape, and not related to head-rotation movements (Extended Data Fig. 4k), which indicates that dmSC neurons encode a variable that is correlated with the likelihood of escape. In agreement with the threat-stimulus data, dPAG neurons are active only during, and not before, escape initiation (Fig. 2k). In addition, there was a correlation between escape speed and peak calcium activity, which was approximately three times stronger in the dPAG than in the dmSC, and was specific for running during escape to the shelter (Extended Data Fig. 4l, m).

These activity profiles are consistent with dmSC neurons representing a pre-escape variable, such as threat intensity, whereas dPAG neurons encode the result of the threat-threshold computation. This predicts that direct activation of the dmSC should produce psychometric and chronometric curves that are similar to those produced by sensory stimulation, as activity is still being passed through the threshold mechanism to initiate escape, whereas dPAG stimulation should reliably elicit escape behaviour with short reaction times. We tested this prediction using in vivo channelrhodopsin-2 (ChR2) activation of dmSC or dPAG VGluT2$^+$ neurons (Fig. 3a), which recapitulated shelter-directed flights (Extended Data Fig. 6a–c, Supplementary Video 5). Gradually increasing the activation of the dmSC network by increasing light intensity progressively increased the escape probability and decreased the response variability (Fig. 3b, c), whereas increasing dPAG activity produced a steep, all-or-none curve, with stereotyped responses for each intensity (Fig. 3b, c), in agreement with our model hypothesis. Reaction times also decreased with stronger dmSC activation, whereas escape latencies for dPAG activation were short across the stimulation range (Fig. 3d), demonstrating that dmSC activity determines the escape onset. Stimulation strength was also correlated with escape speed, but the correlation was stronger for dPAG than for dmSC stimulation (Fig. 3e), which suggests that dPAG activity represents a post-threshold variable from which escape vigour is computed. Moreover, dmSC activation while inactivating the dPAG did not elicit escape, whereas inactivation of an alternative dmSC projection target—the parabigeminal nucleus (PBGN)[5]—did not impair escape, suggesting that threat information from the dmSC has to flow through the dPAG to initiate escape (Extended Data Fig. 6d–i).

To determine whether mSC neurons project directly to dPAG neurons, we performed monosynaptic rabies tracing. This revealed a feed-forward connection with a 11:1 SC:dPAG convergence ratio, consisting of mostly medially located excitatory cells (Fig. 4a; Extended Data Fig. 7). Optogenetic activation of VGluT2$^+$ dmSC axons in vitro elicited excitatory monosynaptic input in 61% of VGluT2$^+$ dPAG neurons (Fig. 4b, left; Extended Data Fig. 8a–e). However, the connections were weak (peak excitatory postsynaptic current (EPSC): −37.9 ± 11.9 pA), with high failure rates (20.3 ± 8%) and low quantal content (2.3 ± 0.6), and followed Poisson statistics, indicating a very low synaptic release probability (Fig. 4c, Extended Data Fig. 8f–h). Consequently, the probability of firing dPAG neurons was extremely low (0.02 ± 0.01 for single light-pulses; Fig. 4d, e), providing a synaptic threshold for dmSC activity to engage the dPAG. However, repeated light stimulation elicited more action potentials than would be expected from temporal summation (spikes per pulse: 0.17 ± 0.1 for 10 Hz, 0.16 ± 0.08 for 20 Hz; membrane time constant = 28.3 ± 3 ms, significantly different from the 20-Hz inter-stimulus interval, $P = 5.8 \times 10^{-6}$, one-sample t-test against 50 ms; Fig. 4e and Extended Data Fig. 8b). This happens because first, the connection facilitates (20 Hz paired-pulse ratio (PPR) = 1.22 ± 0.09, 10 Hz PPR = 1.04 ± 0.08), which provides input amplification at the synaptic level (Fig. 4f). Second, dmSC stimulation triggered a long-lasting increase in the frequency of spontaneous EPSCs (sEPSCs), which decayed to baseline with a time constant of 0.49 s (Fig. 4g). Recordings of VGluT2$^+$ dPAG–dPAG and dmSC–dmSC connectivity showed weak and sparse dPAG input onto dPAG cells (27%, −54 ± 8.3 pA), whereas 100% of dmSC cells received strong input from other dmSC cells (−146.7 ± 41.5 pA, Fig. 4h), which is in agreement with previous work[21] and suggests that recurrent excitation in the dmSC amplifies signals at the network level. Together, these synaptic and network mechanisms allow sustained dmSC activation to overcome the weak connection to VGluT2$^+$ dPAG neurons and drive firing of the escape network. In vivo silicon probe recordings in awake, head-fixed mice showed that during threat stimuli[22,23], dmSC single units fire in the short-term facilitation frequency range of the dmSC–dPAG synaptic connection (73 units from 3 mice, Extended Data Fig. 9), in a contrast-dependent manner (peak firing rate: 20.4 ± 4.1 Hz for 98%, 10.7 ± 1.8 Hz for 50%, 23.9 ± 2.5 Hz for sound, Fig. 4i). Moreover, a fraction of units sustained increased firing beyond the stimulus (37% of visual- and 15% of sound-responding units; time constant to decrease to baseline: 0.23 s and 5.8 s, respectively; Fig. 4j), in agreement with recurrent dmSC activity assisting with the integration to threshold. In the final set of experiments, we tested whether the dmSC–dPAG connection is critical for computing escape. We co-expressed the synaptically-targeted inhibitory designer receptor hM4D-neurexin (hM4D^{nrxn})[24] and ChR2 in VGluT2$^+$ dmSC neurons, which caused a 71 ± 7% reduction in synaptic transmission to the dPAG in the presence of clozapine-N-oxide (CNO), while leaving dmSC neuron firing intact (Fig. 4k, Extended Data Fig. 10a, b). In vivo microinfusion of CNO over dmSC–dPAG

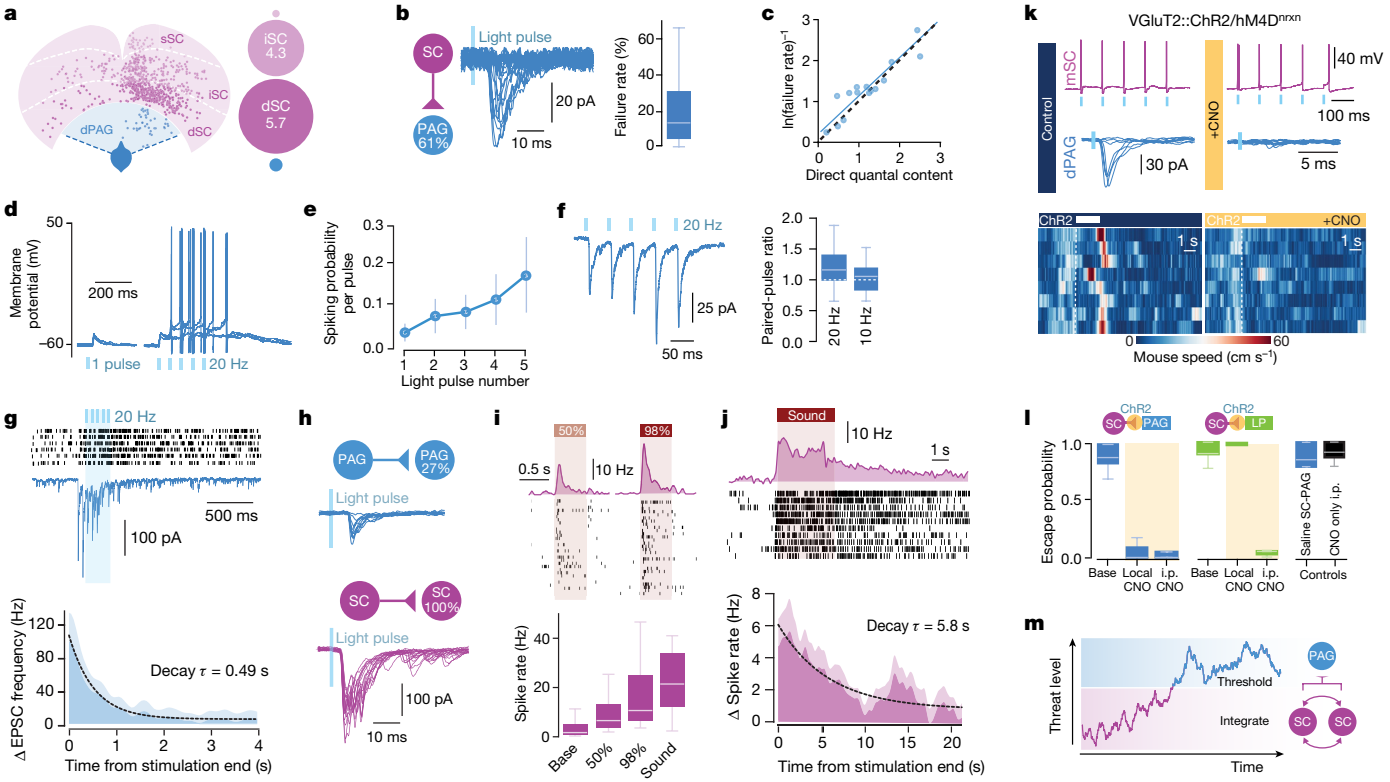


Fig. 4 | Neural circuit and biophysical mechanisms for computing escape behaviour. **a**, Left, dPAG VGluT2$^+$ (blue) and presynaptic cells (pink) from rabies tracing, for deep SC (dSC), intermediate SC (iSC) and superficial SC (sSC). Right, SC:dPAG convergence ratios for single dPAG cells. **b**, mSC–dPAG connectivity (left; $n = 79$ cells, $n = 21$ mice), example traces (middle), and failures summary (right, $n = 8$ cells, $n = 7$ mice). **c**, Direct quantal content versus estimation from failure rate (fit slope = 0.92, 95% confidence interval (0.74, 1.1); $n = 15$ cells). Blue line, linear fit; dashes, unity line (see Methods). **d**, dPAG voltage response to mSC stimulation. **e**, Spiking probability summary ($n = 20$ cells, $n = 10$ mice; plot shows mean and s.e.m.). **f**, Example average trace (left) and summary (right; $n = 11$, 18 cells; $n = 7$, 8 mice for 10 Hz, 20 Hz respectively). **g**, dPAG example trace during mSC stimulation (middle) and sEPSC raster (5 trials, top). Bottom, sEPSC frequency summary ($n = 21$ cells, $n = 9$ mice). Dashed line, exponential fit. **h**, Examples and connectivity for dPAG (top; $n = 11$ cells, $n = 2$ mice) and dmSC (bottom; $n = 22$ cells, $n = 10$ mice). **i**, Top, firing-rate histograms and spike rasters from one dmSC single unit. Bottom, summary data (32 visual- and 45 sound-responsive units, $P = 0.01$ for 50% versus 98%, $P = 2.8 \times 10^{-5}$ for 50% visual versus sound). **j**, Top, example unit showing persistent activity, and average histogram for cells with persistent activity (bottom, dashed line shows exponential fit; 7 units). **k**, hM4D^{nrxn} activation does not affect ChR2-evoked mSC cell firing (top), but blocks mSC–dPAG EPSCs (middle). Bottom, example speed rasters during mSC activation before (left) and after (right) CNO microinfusion to the mSC–dPAG projection. **l**, Summary of CNO application to mSC-PAG (local CNO injection compared to baseline: $P_{escape} = 0.08 \pm 0.05$ versus $P_{escape\text{-}base} = 0.87 \pm 0.03$, $n = 5$ mice, $P = 0.0008$; i.p. CNO compared to baseline: $P_{escape} = 0.05 \pm 0.05$ versus $P_{escape\text{-}base} = 0.97 \pm 0.03$, $n = 4$ mice, $P = 0.01$; $P = 0.5$ for local versus i.p. CNO), and mSC–LP projection (local CNO compared to baseline: $P_{escape} = 1.0 \pm 0$ versus $P_{escape\text{-}base} = 0.95 \pm 0.03$, $n = 4$ mice, $P = 0.1$; i.p. CNO versus baseline: $P_{escape} = 0.04 \pm 0.02$ versus $P_{escape\text{-}base} = 0.89 \pm 0.06$, $n = 3$ mice, $P = 0.04$; $P = 0.01$ for local versus i.p. CNO). Saline mSC–dPAG microinfusion and CNO i.p. injection without hM4D^{nrxn} do not reduce escape ($n = 5$ mice, $P > 0.15$). **m**, Escape decision model. Shaded areas show s.e.m; box-and-whisker plots show median, IQR and range. *P* values: two-tailed *U*-test.

synapses blocked escape in response to visual stimuli (Extended Data Fig. 10c) and optogenetic dmSC activation, similar to systemic CNO injection (Fig. 4k, l, Supplementary Video 6). Notably, doubling the intensity or the frequency of optogenetic stimulation was not sufficient to rescue escape (Extended Data Fig. 10a, d), whereas inhibiting the dmSC projection to the lateral posterior nucleus of the thalamus (LP) did not affect escape (Fig. 4l).

Our results support a model in which threat evidence is integrated in the dmSC and passed through a synaptic threshold at the dPAG level to initiate escape (Fig. 4m). Although it is likely that several mSC projections support escape behaviour, we show that the dmSC–dPAG synaptic connection is required for the initiation of escape, whereas SC projections to LP[5,7] are not, which suggests that there might be dedicated projections for controlling freezing[7] and escape. Also, in contrast to previous work[5] using optogenetic activation of SC projections to the PBGN, we did not find a critical role for this pathway in escape initiation, which could be explained in previous studies by antidromic activation of SC neurons projecting to both PBGN and dPAG, or by back-projections to the SC. A key result is that dmSC activity encodes a high-order signal predictive of escape, in agreement with its role in multisensory integration[25] and decision making[26–28]. Successfully escaping from threats to reach safety requires the integration of multiple information streams, including knowledge about the spatial environment[9], and our results provide a mechanistic entry point for understanding how the brain computes a fundamental survival behaviour, and goal-directed behaviours in general.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0244-6.

Received: 21 June 2017; Accepted: 11 May 2018;
Published online: 20 June 2018

1. Ydenberg, R. C. & Dill, L. M. The economics of fleeing from predators. *Adv. Study Behav.* **16**, 229–249 (1986).
2. De Franceschi, G., Vivattanasarn, T., Saleem, A. B. & Solomon, S. G. Vision guides selection of freeze or flight defense strategies in mice. *Curr. Biol.* **26**, 2150–2154 (2016).
3. Yilmaz, M. & Meister, M. Rapid innate defensive responses of mice to looming visual stimuli. *Curr. Biol.* **23**, 2011–2015 (2013).

4. Kunwar, P. S. et al. Ventromedial hypothalamic neurons control a defensive emotion state. *eLife* **4**, e06633 (2015).
5. Shang, C. et al. A parvalbumin-positive excitatory visual pathway to trigger fear responses in mice. *Science* **348**, 1472–1477 (2015).
6. Wang, L., Chen, I. Z. & Lin, D. Collateral pathways from the ventromedial hypothalamus mediate defensive behaviors. *Neuron* **85**, 1344–1358 (2015).
7. Wei, P. et al. Processing of visually evoked innate fear by a non-canonical thalamic pathway. *Nat. Commun.* **6**, 6756 (2015).
8. Xiong, X. R. et al. Auditory cortex controls sound-driven innate defense behaviour through corticofugal projections to inferior colliculus. *Nat. Commun.* **6**, 7224 (2015).
9. Vale, R., Evans, D. A. & Branco, T. Rapid spatial learning controls instinctive defensive behavior in mice. *Curr. Biol.* **27**, 1342–1349 (2017).
10. Gross, C. T. & Canteras, N. S. The many paths to fear. *Nat. Rev. Neurosci.* **13**, 651–658 (2012).
11. Blanchard, R. J., Blanchard, D. C., Rodgers, J. & Weiss, S. M. The characterization and modelling of antipredator defensive behavior. *Neurosci. Biobehav. Rev.* **14**, 463–472 (1990).
12. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
13. Carandini, M. & Churchland, A. K. Probing perceptual decisions in rodents. *Nat. Neurosci.* **16**, 824–831 (2013).
14. Dean, P., Redgrave, P. & Westby, G. W. M. Event or emergency? Two response systems in the mammalian superior colliculus. *Trends Neurosci.* **12**, 137–147 (1989).
15. Deng, H., Xiao, X. & Wang, Z. Periaqueductal gray neuronal activities underlie different aspects of defensive behaviors. *J. Neurosci.* **36**, 7580–7588 (2016).
16. Fotowat, H. & Gabbiani, F. Collision detection as a model for sensory-motor integration. *Annu. Rev. Neurosci.* **34**, 1–19 (2011).
17. Shea-Brown, E., Gilzenrat, M. S. & Cohen, J. D. Optimization of decision making in multilayer networks: the role of locus coeruleus. *Neural Comput.* **20**, 2863–2894 (2008).
18. Silva, B. A. et al. Independent hypothalamic circuits for social and predator fear. *Nat. Neurosci.* **16**, 1731–1733 (2013).
19. Tovote, P. et al. Midbrain circuits for defensive behaviour. *Nature* **534**, 206–212 (2016).
20. Wietek, J. et al. An improved chloride-conducting channelrhodopsin for light-induced inhibition of neuronal activity *in vivo*. *Sci. Rep.* **5**, 14807 (2015).
21. Pettit, D. L., Helms, M. C., Lee, P., Augustine, G. J. & Hall, W. C. Local excitatory circuits in the intermediate gray layer of the superior colliculus. *J. Neurophysiol.* **81**, 1424–1427 (1999).
22. Gale, S. D. & Murphy, G. J. Active dendritic properties and local inhibitory input enable selectivity for object motion in mouse superior colliculus neurons. *J. Neurosci.* **36**, 9111–9123 (2016).
23. Zhao, X., Liu, M. & Cang, J. Visual cortex modulates the magnitude but not the selectivity of looming-evoked responses in the superior colliculus of awake mice. *Neuron* **84**, 202–213 (2014).
24. Stachniak, T. J., Ghosh, A. & Sternson, S. M. Chemogenetic synaptic silencing of neural circuits localizes a hypothalamus→midbrain pathway for feeding behavior. *Neuron* **82**, 797–808 (2014).
25. Schiller, P. H. in *The Handbook of Physiology* Vol. 3 (eds Brookhart, J. M. and Mountcastle, V. B.) 457–505 (Lippincott Williams and Wilkins, Pennsylvania, 1984).
26. Felsen, G. & Mainen, Z. F. Midbrain contributions to sensorimotor decision making. *J. Neurophysiol.* **108**, 135–147 (2012).
27. Cohen, J. D. & Castro-Alamancos, M. A. Neural correlates of active avoidance behavior in superior colliculus. *J. Neurosci.* **30**, 8502–8511 (2010).
28. Horwitz, G. D., Batista, A. P. & Newsome, W. T. Representation of an abstract perceptual decision in macaque superior colliculus. *J. Neurophysiol.* **91**, 2281–2296 (2004).

Acknowledgements This work was funded by a Wellcome Trust Henry Dale Fellowship (098400/Z/12/Z), Medical Research Council (MRC) grant MC-UP-1201/1, Wellcome Trust/Gatsby Charitable Foundation SWC Fellowship (T.B.), MRC PhD Studentship (D.A.E., R.V.), Boehringer Ingelheim Fonds PhD fellowship (R.V.), DFG fellowship (A.V.S., S.R.), Marie Skłodowska-Curie Individual Fellowship (706136) and EMBO Long Term Fellowship (Y.L.). We thank P. Latham and members of the Branco laboratory for discussions; S. Sternson, P. Dayan, T. Margrie and T. Mrsic-Flogel for comments on the manuscript; S. Sternson, S. Wiegert, T. Oertner and T. Margrie for gifts of viral vectors; P. Iordanidou, T. Okbinoglu, L. Jin, the LMB and SWC Biological Research Facility and FabLabs for technical support; D. Campagner, T. Harris and N. Steinmetz for help with silicon probe recordings; and K. Betsios for programming the data acquisition software.

Reviewer information *Nature* thanks V. Bolshakov, P. Tovote and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.B. and D.A.E. conceived the project with input from A.V.S., R.V. and S.R.; D.A.E., A.V.S., R.V. and S.R. performed behavioural and optogenetic experiments. T.B. performed theoretical modelling, D.A.E. performed calcium imaging, A.V.S. and T.B. performed in vitro electrophysiology, Y.L. and R.V. performed single-unit recordings, D.A.E. and A.V.S. performed chemogenetic experiments, A.V.S. performed anatomical tracing. All authors analysed data and contributed to the experimental design. T.B. supervised the project. T.B. wrote the manuscript with help from D.E. and A.V.S.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0244-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0244-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mice. Male and female adult C57BL/6J wild-type, VGluT2-ires-Cre[29] (Jackson Laboratory, stock 016963) and VGluT2::eYFP (R26 eYFP, Jackson Laboratory 006148; eYFP, enhanced yellow fluorescent protein) mice were housed with free access to chow and water on a 12:12 h light:dark cycle and tested during the light phase. All experiments were performed under the UK Animals (Scientific Procedures) Act of 1986 (PPL 70/7652) following local ethical approval. Minimum sample sizes were predetermined from power estimates based on pilot experiments. Animals in test and control groups were littermates and randomly selected. Behavioural experiments were not performed blinded as the experimental setup is closed-loop and automatically delivers stimuli. Behavioural data were annotated blinded and by several experimenters.

Surgical procedures. Mice were anaesthetized with an intraperitoneal (i.p.) injection of ketamine (95 mg kg^{-1}) and xylazine (15.2 mg kg^{-1}), and carpofen (5 mg kg^{-1}) was administered subcutaneously. Isoflurane (0.5–2.5% in oxygen, 1 l min^{-1}) was used to maintain anaesthesia. Craniotomies were made using a 0.5-mm burr and viral vectors were delivered using pulled glass pipettes (10 µl Wiretrol II with a Sutter P-1000) in an injection system coupled to a hydraulic micromanipulator (MO-10, Narishige) on a stereotaxic frame (Model 1900 and 963, Kopf Instruments), at approximately 10 nl min^{-1}. Implants were affixed using light-cured dental cement (RelyX Unicem 2, 3M) and the wound sutured (6-0, Vicryl Rapide) or glued (Vetbond). Coordinates are measured from lambda.

Viruses. The following viruses were used in this study and are referred to by contractions in the text. For optogenetic activation, adeno-associated virus (AAV) AAV2-EF1a-DIO-hChR2(H134R)-eYFP-WPRE (3.9×10^{12} genome copies per ml (GC ml^{-1})) and AAV2-EF1a-DIO-hChR2(H134R)-mCherry-WPRE (6.6×10^{12} GC ml^{-1}; Deisseroth) were acquired from the UNC Vector Core. Optogenetic inhibition experiments were performed with AAV9-Ef1a-DIO-iChlo-2A-tDimer (3.75×10^{12} GC ml^{-1}; a gift from S. Wiegert and T. Oertner) or AAV1-EF1a-DIO-iChloC-2A-dsRed (5×10^{13} GC ml^{-1}; Addgene 70762, a gift from T. Margrie). For control and calcium-imaging experiments respectively, AAV2-EF1a-DIO–eYFP-WPRE (4.0×10^{12} GC ml^{-1}) and AAV9-CAG-DIO-GCaMP6s-WPRE (6.25×10^{12} GC ml^{-1}) were acquired from Penn Vector Core. For retrograde rabies tracing, EnvA pseudotyped SADB19 rabies virus (EnvA-dG-RV-mCherry) was used in combination with AAV8 coding for the EnvA receptor TVA and rabies virus glycoprotein (RG) that were prepared from pAAV-EF1a-FLEX-GT (Addgene plasmid 26198, Callaway) and pAAV-Syn-Flex-RG-Cerulean (Addgene plasmid 98221, Margrie). All viruses used for rabies tracing were a gift from T. Margie[30], and had been previously tested for leakiness and specificity[31]. Additionally, a recombinant AAV with retrograde functionality (rAAV2-retro-mCherry, 6.97×10^{12} GC ml^{-1}, Addgene 81070[32]) was used. For chemogenetic inhibition experiments, AAV5-CAG-DIO-mCherry-2A-hM4D-HA-2A-nrxn1A (3.9×10^{12} GC ml^{-1}, a gift from S. Sternson) or AAV2-CAG-DIO-mCherry-2A-hM4D-nrxn1A (6.19×10^{11} GC ml^{-1}, Addgene 60544) were used.

Behavioural procedures. *Experimental set-up.* All behavioural experiments were performed in a rectangular Perspex arena (W: 20 cm × L: 60 cm × H: 40 cm) with a red-tinted shelter (19 cm × 10 cm × 13.5 cm) at one end, housed within a sound-deadening, light-proofed cabinet with six infrared light-emitting diode (LED) illuminators (TV6700, Abus). A screen (90 cm × 70 cm; '100 micron drafting film', Elmstock) was suspended 64 cm above the arena floor, and a DLP projector (IN3126, InFocus) back-projected a grey uniform background via a mirror, providing 7–8 lx at the arena floor. Experiments were recorded at 50 frames per second with a near-IR GigE camera (acA1300-60gmNIR, Basler) positioned above the arena centre. Video recording, sensory and optogenetic stimulation was controlled with custom software written in LabVIEW (2015 64-bit, National Instruments) and Mantis software (mantis64.com). The position of the mouse was tracked online, and used to deliver stimuli when the mouse entered a predefined 'threat area' (21 cm × 20 cm area at opposite end to shelter). An empty plastic Petri dish (replaced fresh for each mouse; 35 mm) was affixed to the arena floor in the centre of the threat area to enrich the environment. All signals and stimuli, including each camera frame, were triggered and synchronised using hardware-time signals controlled with a PCIe-6351 board (National Instruments).

Protocols. Mice were placed in the arena and given 8 min to explore the new environment, after which sensory stimuli were delivered when the mouse entered the threat area for longer than 100 ms. A typical experiment lasted 30–60 min. In the standard visual stimulation protocol, we used a pseudo-random contrast sequence to minimise the development of aversion or habituation during the behavioural session (see Extended Data Figs. 1 and 5e, f for quantification). The sequence consisted of a first stimulus at 98% contrast, followed by a random selection without replacement from the remaining contrasts, and this process was repeated until the end of the behavioural session. Each stimulus was delivered with an inter-stimulus interval of at least 30 s. For the conditioning protocol shown in Fig. 2k and Extended Data Fig. 5, repeated presentations (3–6 trials) at 98% contrast were delivered with no minimum inter-stimulus interval after a 10-min acclimatization period.

Sensory stimuli. The standard visual stimulus was a sequence of five dark expanding circles, and unless otherwise stated, each subtended a visual angle of 2.6° at onset and expanded linearly at 118° s^{-1} to 47° over 380 ms, after which it maintained the same size for 250 ms and began an inter-stimulus interval of 500 ms. The contrast of the spot was varied in a number of experiments, and for clarity is reported as a positive percentage (low to high; for example, 25% to 98%), converted from the negative Weber fraction (low to high; −0.25 to −0.98). The contrast was varied by altering the intensity of the spot against a grey screen maintained at constant luminance (standard luminance, 7.95 cd m^{-2}). The spot was located on the screen directly above the centre of the threat area, approximately 15° from the zenith of the mouse. The auditory stimulus consisted of a frequency-modulated upsweep from 17 to 20 kHz over 3 s (ref. [33]). Waveform files were created in MATLAB (Mathworks), and the sound was generated in LabVIEW, amplified and delivered via an ultrasound speaker (L60, Pettersson) positioned 50–55 cm above the arena, centred over the threat area.

Analysis. Behavioural video and tracking data was sorted into peri-stimulus trials and manually annotated. Detection of the threat stimulus was assumed if the mouse showed a stimulus-detection response, in which the ears of the mouse move posteriorly and ventrally, which precedes interruption or commencement of body movement. To differentiate failures of escaping from failures of attending to the stimulus, trials with no stimulus-detection response were excluded from the analysis. This resulted in the exclusion of three no-escape trials from the 25% contrast dataset, which increased the escape probability from 0.12 to 0.13. The onset of escape was measured as the first video frame marking the onset of a continuous movement consisting of a head turn followed by a full-body turn towards the shelter. Escape was annotated automatically and defined as the mouse moving to enter the shelter in a single movement without stopping, within 0.9 s after stimulus termination (or 6 s after approaching a 15-cm boundary from the threat area for spontaneous escapes after conditioning). Behaviour metrics were calculated by pooling all trials and mice (Fig. 1d–f) and also by analysing each mouse individually and then computing an average value across all mice (Extended Data Fig. 1a–c). Statistical analysis was performed using the number of mice as the sample size. The escape probability for a given stimulus is the fraction of trials which led to an escape to the shelter. The maximum speed of the escape is calculated as the peak value of the speed trace between the onset of the escape and entry to the nest. Quantification of exploratory behaviour was done for behavioural sessions lasting at least 40 min, by calculating the cumulative displacement of the mouse in 1-min bins followed by smoothing with a five-point flat window. We did not observe any differences in the behavioural response to threat stimulation between male and female mice, and therefore data from both sexes has been pooled (for 98% contrast stimulation, escape probability: 0.86 for males, 0.88 for females, $P = 1.0$, Fisher exact test; reaction time: 369.2 ± 51.8 ms for males, 365.6 ± 39.6 ms for females, $P = 0.96$, two-tailed t-test; vigour: 91.8 ± 4.5 cm s^{-1} for male, 89.1 ± 11.1 for female, $P = 0.81$, two-tailed t-test).

Behavioural model. The threat level (T) evolves over time according to

$$\tau_T \frac{dT}{dt} = -T + Ca(t) + \sigma_N W$$

where $a(t)$ is the diameter of the expanding visual spot scaled by the spot contrast C. The variable τ_T sets the time constant for changing the threat level and W is a white-noise Wiener process parametrised by σ_N. At each time point, T is compared against a threshold B, and escape initiated if $T > B$. The reaction time is the time at threshold crossing measured relative to stimulus onset. In this model we allow the threat level to continue evolving after the threshold has been crossed, similar to previous work on changes of mind during decision making[34], and escape vigour V is computed from the peak of the threat level as a logistic function:

$$V = \frac{1}{1 + e^{(-k(T - Bs))}}$$

The model was first fitted with three free parameters (B, τ_T, σ_N) to the reaction time and escape probability data simultaneously by simulating 10,000 trials for each parameter set and using the brute force method in LMFIT Python 2.7 package. Escape vigour was then fitted to the average peak threat levels across all trials with free parameters k and s using least-squares minimisation in LMFIT. The fit parameters for the curves shown in Fig. 1 are: $B = 0.165$, $\tau_T = 1{,}200$ ms, $\sigma_N = 0.6$, $k = 90$, $s = 1.5$.

Pharmacological inactivation. Mice were bilaterally implanted with guide cannulae (Plastics One, Bilaney Consultants) over the target region (see Supplementary Table 1) and given at least 48 h for recovery. On the test day, mice were placed in the standard arena for 10 min and escape responses were assessed with a single visual stimulus (one 98% contrast expanding spot) or auditory stimulus. Additionally, in PBG- and PAG-cannulated mSC-VGluT2::ChR2 mice, optogenetic responses were also evoked. The mice were then lightly anaesthetized in an induction chamber and placed on a heating pad where anaesthesia was maintained with a nose

cone (2% isoflurane, 1 l min^{-1}). Internal cannulae were inserted and sealed with Kwik-Sil. Muscimol-BODIPY-TMR-X (0.5 mg ml^{-1}) or Alexa-555 (100 μM; Life Technologies), dissolved in 1:1 phosphate-buffered saline (PBS): 0.9% saline with 1% dimethyl sulfoxide (DMSO), was then infused at a rate of 70–100 nl min^{-1} using a microinjection unit (10 μl Model 1701 syringe; Hamilton, in unit Model 5000; Kopf Instruments) followed by a 5-min wait period per hemisphere. Mice spent no longer than 30 min under anaesthesia and were given 30 min to recover in the home cage, after which they were placed back in the cleaned arena and subjected to visual, auditory or optogenetic stimulation. Immediately upon termination of the behavioural assay, around 1 h after infusion, mice were anaesthetized with isoflurane (5%, 2 l min^{-1}) and decapitated. Acute slices (150 μm) were cut using a microtome (Campden 7000smz-2 or Leica VT1200S) in ice-cold PBS (0.1 M), directly transferred to 4% paraformaldehyde (PFA) solution, and kept for 20 min at 4 °C. The slices were then rinsed in PBS, counter-stained with 4′,6-diamidino-2-phenylindole (DAPI; 3 μM in PBS), and mounted on slides in SlowFade Gold (Life Technologies) before wide-field imaging (Nikon TE2000) on the same day to confirm the site of infusion. Behavioural data was annotated as described. For the calculation of the maximum exploration speed, the peak speed of the 7-min acclimatisation period before stimulation was used. Statistical analysis was performed using the number of mice as the sample size.

Calcium imaging in freely-moving mice. *Data acquisition.* A miniaturised head-mounted fluorescence microscope[35] (Model L, Doric Lenses Inc.) was used to image GCaMP6s in neurons of male VGluT2-Cre mice. AAV9-CAG-Flex-GCaMP6s (300–550 nl; Penn Vector Core) was injected into the mSC (anteroposterior, AP: −0.2 to −0.5; mediolateral, ML: +0.25; dorsoventral, DV: −1.6) or dPAG (AP: −0.4 to −0.6, ML: +0.25, DV: −2.2). At the level of the inferior colliculus, the dura was incised using a 30G needle, and gently pulled forward to partially reveal the SC. A GRIN lens-equipped cannula (SICL_V_500_80; Doric Lenses Inc.) was used to push forward the transverse sinus and inserted to the same depth as the injection coordinates, after which the craniotomy was covered with Kwik-Cast and the cannula affixed with dental cement. At least 21 days after surgery, the microscope was attached to the mouse without anaesthesia, and the mouse was placed back in the home cage for 5–10 min, for acclimatisation to the microscope. During this period, the optimal imaging parameters for the preparation were determined: the acquisition rate was 14.2 Hz in most experiments (median; range: 10–20 Hz) with an excitation power of 450 μW (median; range: 0.2–1.1 mW). After a baseline period of 7 min, mice were exposed to visual and/or auditory stimulation. For the visual stimulation, the contrast was 98%, the inter-stimulus interval was 750 ms, and the post expansion period was 20 ms, with the total epoch length and expansion rate unchanged. A typical session lasted 1.5 h (1–3 sessions per mouse), with imaging data acquired during stimulation and control trials in approximately 5-min epochs, with at least 2 days between sessions. If the mouse showed prolonged bouts of inactivity, imaging was halted to minimize photobleaching. Fluorescence and behavioural frame trigger signals were acquired at 10 kHz for offline synchronisation.

Data analysis. Behavioural video and tracking data were sorted into peri-stimulus trials and manually annotated to mark behavioural events as described above. Fluorescence stacks were registered[36] and background-subtracted (Fiji). Cell body-like structures were identified manually as regions-of-interest (ROIs; elliptic or polygonal areas) in Fiji using the maximum intensity projection of registered movies, aided by inspection of deconvolved images. For each mouse, ROI masks were rigidly translated to account for field-of-view (FOV) movement between imaging sessions, and new cells added to the FOV if they became visible. In some cases, the FOV moved such that ROIs could not be mapped to the previous sessions, and it was therefore counted as a new FOV. Mean intensity traces were extracted for each ROI, interpolated with the behavioural video frames and tracking data, and the change in fluorescence intensity relative to the resting fluorescence intensity ($\Delta F/F$) calculated on a trial-by-trial basis with a baseline of 5 s before stimulus onset. Traces were then smoothed with a 20-point Hanning window and z-scored. ROIs were only included in the analysis if they had transients with a z-score above 2 at any time during the recording session, to ensure that they were live, active neurons. Average responses for each cell were obtained by averaging across all trials independent of the trial outcome and statistical analysis was performed on all cells pooled together. Ensemble average responses were obtained by averaging the responses of all cells in a FOV and summary statistics calculated over all trials for each FOV. For the ROC analysis, the annotated behavioural outcomes were used to sort data into 'Escape' and 'No Escape' classes, and the ROC curves and AUC statistics were calculated using the open-source package Scikit-learn. The s.d. for the AUC was estimated using bootstrapping. 'Peri' and 'Pre-escape' time periods were defined as escape onset ± 1 s and <1 s, respectively. For the plot in Extended Data Fig. 4i, escape latencies were first binned and average calcium signal waveforms calculated for each bin, and the signal rise slope was obtained by fitting a linear function ($y = mx + b$). The onset of calcium signals was measured by finding the time of the peak and iteratively moving backwards along the signal to determine the time point at which the signal reaches the baseline. Peak calcium responses after conditioning were taken from a 5-s time window starting when the mouse entered the threat area.

Optogenetic experiments. For optogenetic activation[37], VGluT2-Cre and VGluT2::eYFP mice were injected with AAV-DIO-ChR2-eYFP or -mCherry, (see Viruses) into the dmSC (80–120 nl per side, ML: ±0.2 to 0.35, AP: −0.25 to −0.45, DV: −1.4 to −1.55) or dPAG (40–80 nl per side ML: ±0.0 to −0.4, AP: −0.4 to −0.6, DV: −1.95 to −2.2). Control mice were injected with 120 nl AAV2-DIO–eYFP into the dPAG. One optic fibre (200-μm diameter, MFC-SMR; Doric Lenses Inc.) was implanted per mouse, medially, 250–300 μm dorsal to the injection site. For optical stimulation, light was delivered by a 473-nm solid-state laser (CNI) in conjunction with a continuous neutral density filter wheel for varying light intensity (NDC-50C-4M, Thorlabs) and a shutter (LS6, Uniblitz) driven by trains of pulses generated in LabVIEW. In some experiments, this system was substituted by a laser diode module (Stradus, Vortran) with direct analogue modulation of laser intensity. Magnetic patchcords (Doric Lenses Inc.) were combined with a rotary joint (FRJ 1×1, Doric Lenses Inc.) to allow the cannula to be connected without restraint and allow unhindered movement. In all experiments, mice were placed in the standard arena and given 8 min to acclimatise. As the fraction of cells spiking in a ChR2-expressing neuronal network increases as a function of light intensity in vivo[38], we chose to systematically modulate light intensity as a proxy for setting the level of activation in the dPAG and mSC. For the intensity modulation assay, the laser intensity was set initially to give a low irradiance (0.1–0.2 mW mm^{-2}) that did not evoke an observable behavioural response. Mice were photostimulated (473 nm, train of 10 light pulses of 10 ms at 10 Hz) upon entering the threat area with an inter-stimulus interval of at least 30 s. After at least three trials of this intensity, the irradiance was increased by 0.1–0.3 mW mm^{-2} until a behavioural response was observed, after which 8–15 trials were obtained at a given intensity, before further increasing the light intensity. This process was iterated until an intensity was reached which always evoked a flight response ($P_{escape} = 1$). For one mouse, the standard stimulus was not sufficient to reach $P_{escape} = 1$ and the curve was acquired with a higher frequency stimulus (10 light pulses of 10 ms at 20 Hz). If the mouse stopped exploring the arena, precluding $P_{escape} = 1$ from being obtained, the experiment was terminated after 4 h and not analysed. To normalize stimulation intensity and compare across mice, trials were first classified as escape if the mouse reached the shelter within 5 s of stimulation onset, to calculate the fraction of escape trials at a given intensity. The escape probability curve of each mouse was then fitted with a logistic function ($1/(1 + e^{-k(x-x0)})$), and light intensities were normalized to $x0$. In the frequency modulation assay, high laser power was used (range, 12–13.5 mW mm^{-2}) and the stimulus consisted of 10 light pulses of 10 ms at either 2, 5, 10, 20 and 40 Hz, delivered in a pseudo-random order.

For histological confirmation of the injection site, mice were anaesthetized with Euthatal (0.15–0.2 ml) and transcardially perfused with 10 ml of ice-cold PBS with heparin (0.02 mg ml^{-1}) followed by 4% paraformaldehyde (PFA) in PBS solution. Brains were post-fixed overnight at 4 °C then transferred to 30% sucrose solution for 48 h. Sections (30 μm) were cut with a cryostat (Leica CM3050S) and a standard free-floating immunohistochemical protocol was used to increase the signal of the tagged ChR2 and counter-stain neurons. The primary antibodies used were anti-GFP (1:1,000, chicken; A10262, or rabbit; A11122, Life Technologies), anti-RFP (1:1,000, rabbit; 600-401-379, Rockland) and anti-NeuN (1:1,000, mouse; MAB-377, Millipore) and the secondary antibodies were Alexa-488 Donkey anti-rabbit and Goat anti-chicken, Alexa-568 Donkey anti-rabbit and Donkey anti-mouse, and Alexa-647 Donkey anti-mouse (1:1,000, Life Technologies). Brain sections were mounted on charged slides using the mounting medium SlowFade Gold (containing DAPI; S36938, Life Technologies), and imaged using a wide-field microscope (Nikon TE2000).

For optogenetic inactivation experiments, VGluT2-Cre and VGluT2::eYFP mice were injected with AAV-DIO-iChloC-dsRed, (see 'Viruses') into the dmSC (250 nl per side, ML: ±0.35, AP: 0.1 to −0.45, DV: −1.4 to −1.55) or dPAG (200 nl per side, ML: ±0.4, AP: −0.4 to −1, DV: −2.2), with two injections per hemisphere along the AP axis spaced 300 μm apart. Dual optic fibres (400 μm diameter, 1.2 mm apart, DFC_400/430-0.48_3.5mm_GS1.2_C60; Doric Lenses Inc.) were implanted at the injection site. Behavioural testing was done 10–41days after virus injection. Mice were presented with visual or auditory stimuli that elicited escape, and laser-on trials were interleaved with laser-off trials (473 nm, 5–8 s square pulse, 15 mW mm^{-2}). For histological confirmation of the fibre placement and injection site, mice were decapitated under anaesthesia, brains were quickly removed and post-fixed in 4% PFA overnight at 4 °C. Slices of 100 μm thickness were cut on a HM650V vibratome (Microm) in 0.1 M PBS, stained with DAPI before mounting, and imaged on a wide-field microscope (Axio Imager 2, Zeiss).

Chemogenetic inactivation experiments. VGluT2-Cre and VGluT2::eYFP mice were injected with AAV-DIO-hM4D-nrxn-mCherry (see 'Viruses') into the dmSC (200–250 nl per side, ML: ±0.35, AP: −0.1 to −0.45, DV: −1.4 to −1.55), with

2–3 injections per hemisphere along the AP axis. Dual guide cannulae were implanted at ML: ±0.6, AP: −0.55, DV: −1.6 to target the SC–dPAG projection, and ML: ±1.7, AP: +1.7, DV: −2.05 (angle: 7° lateral from zenith) to target the SC–LP thalamus projection. In experiments with optogenetic stimulation, AAV-DIO-ChR2–eYFP was injected into the dmSC first (coordinates and volumes as above) and a 200-μm optic fibre cannula was implanted at ML: ±0.1, AP: −0.3, DV: 1.35 (angle: 35° posterior from zenith). After 20–55 days, escape responses to optogenetic or visual stimuli were assessed in a baseline session to estimate the stimulus intensities that evoke escape with $P_{escape} = 1$. Thirty minutes after microinfusion or i.p. injection, escape responses were reassessed using the same stimuli, and, for optogenetic activation, 200% of baseline intensity or frequency were tested in addition to the baseline strength. For cerebral microinfusions, CNO was diluted in buffered saline containing (in mM): 150 NaCl, 10 D-glucose, 10 HEPES, 2.5 KCl, 1 $MgCl_2$, and to a final concentration of 1 or 10 μM. Experiments with visual-evoked escape were done with 1 μM, and optogenetically-evoked escape with 1 and 10 μM. There was no significant difference between 1 and 10 μM at the electrophysiological and behavioural level, and the data have therefore been pooled (comparisons between 1 μM and 10 μM CNO: ChR2-induced firing of SC VGluT2$^+$ neurons, $P > 0.999$ Wilcoxon test; SC–dPAG VGluT2$^+$ EPSC amplitude, $P = 0.0973$ Mann–Whitney test; P_{escape} after CNO microinfusion, $P = 0.6095$, Mann–Whitney test). Cerebral microinfusions of CNO or vehicle were performed as described above using 500 μm protruding internal cannulae (see Pharmacological Inactivation), with a volume of 0.6–1.0 μl per hemisphere. For i.p. injections, 1 mg CNO was dissolved in 1 ml 0.9% saline just before the experiment and injected at a final concentration of 10 mg kg^{-1}. Repeated administration of CNO was separated by 2–3 days, preceded by a new baseline session for each treatment. Histological confirmation of cannula placements and viral infection was performed as stated above.

Electrophysiological recordings in acute midbrain slices. *Data acquisition.* Coronal slices were prepared from VGluT2::eYFP mice aged 6–12 weeks. Brains were quickly removed and transferred to ice-cold slicing solution containing (in mM): 87 NaCl, 26 $NaHCO_3$, 50 sucrose, 10 glucose, 2.5 KCl, 1.25 NaH_2PO_4, 3 $MgCl_2$, 0.5 $CaCl_2$. Acute coronal slices of 250 μm thickness were prepared at the level of the SC and PAG (−4.8 to −4.1mm from bregma) using a vibratome (VT1200, Leica or 7000smz-2, Campden). Slices were then stored under submerged conditions, at near-physiological temperature (35 °C) for 30 min before being cooled down to room temperature (19–23 °C). For recordings, slices were transferred to a submerged chamber and perfused with artificial cerebrospinal fluid (aCSF) containing (in mM): 119 NaCl, 26 $NaHCO_3$, 10 glucose, 2.5 KCl, 2 $CaCl_2$, 1 $MgCl_2$, 1 NaH_2PO_4 (heated to 34 °C at a rate of 2–3 ml min^{-1}). All aCSF was equilibrated with carbogen (95% O_2, 5% CO_2, final pH 7.3). Whole-cell patch-clamp recordings were performed with an EPC 800 amplifier (HEKA). Data was digitised at 20 kHz (PCI 6035E, National Instruments), filtered at 5 kHz and recorded in LabVIEW using custom software and Mantis software (mantis64.com). Pipettes were pulled from borosilicate glass capillaries (Harvard Apparatus, 1.5-mm OD, 0.85-mm ID) with a micropipette puller (P-1000, Sutter, USA or P-10, Narishige, Japan) to a final resistance of 4–6 MΩ. Pipettes were backfilled with internal solution containing (in mM): 130 potassium gluconate or $KMeSO_3$, 10 KCl, 10 HEPES, 5 phosphocreatine, 2 Mg-ATP, 2 Na-ATP, 1 EGTA, 0.5 Na_2-GTP, 285–290 mOsm, pH was adjusted to 7.3 with KOH. VGluT2$^+$ dPAG and dmSC cells were visualized on an upright Slicescope (Scientifica) using a 60× objective (Olympus) and identified based on location and eYFP expression. The resting membrane potential was determined immediately after establishing the whole-cell configuration and experiments were continued only if cells had a resting membrane potential more hyperpolarized than −45 mV. Input resistance (R_{in}) and series resistance (R_s) were monitored continuously throughout the experiment, and R_s was compensated in current-clamp recordings. Only cells with a stable $R_s < 30$ MΩ were analysed. For ChR2-assisted circuit mapping, recordings were made 10–51 days (mean = 22.3 ± 2.3 days) after injection of AAV2-DIO-ChR2-mCherry into the mSC or dPAG of VGluT2::eYFP mice. ChR2 was stimulated with wide-field 490-nm LED illumination (pe-100, CoolLED, 1-ms or 10-ms pulse length, maximum light intensity = 2.7 mW). To characterize the cellular effects of iChloC activation, dPAG or dmSC VGluT2$^+$ cells expressing AAV5-DIO-iChloC-dsRed were recorded from at 46 days after infection, and iChloC was stimulated with 1-s long, 490-nm light pulses. Recordings in mice expressing hM4D-nrxn in the dmSC were made 22–53 days after injection (mean = 29.4 ± 3.1 days), and ChR2 was activated at 10, 20 and 100% light intensity (0.27, 0.54 and 2.7 mW).

Pharmacology. No drugs were added to the recording aCSF, except for the following experiments: miniature EPSCs (mEPSCs) were recorded in 1 μM tetrodotoxin (TTX, Sigma Aldrich), and ESPC recordings shown in Extended Data Fig. 8 were recorded in 1 μM TTX and 100 μM 4-aminopyridine (4-AP, Sigma Aldrich); to test the effect of hM4D-neurexin activation on firing rates and synaptic transmission, 1–10 μM CNO (free base, Hellobio) was added to the aCSF during recordings.

Data analysis. Analysis was performed using custom-written procedures in Python, except for the analysis of sEPSCs and mEPSCs which was done in IGOR Pro 6 (WaveMetrics) using TaroTools (by Taro Ishikawa). The R_{in} was calculated from the steady-state voltage measured in response to a hyperpolarizing test pulse of 500-ms duration at a holding potential of −60 mV. The membrane time constant was calculated by fitting the decay of the test pulse with a single exponential ($y = y0 + Ae^{(-(x-x0)/\tau)}$). The membrane potential values stated in the text are not corrected for liquid junction potentials. The sEPSC frequency before and after ChR2 stimulation was calculated from 6–8 repetitions per cell. Failures of light-evoked synaptic transmission were defined as a peak amplitude of less than the mean current baseline +2s.d. in a time window defined by the onset of the mean evoked synaptic current ±5 ms. Quantal content calculated by the direct method was obtained by dividing the peak amplitude of the evoked current by the peak amplitude of the sEPSCs in the same cell (which is not significantly different from the mEPSC amplitude, see Extended Data Fig. 8f–h), and the Poisson estimation was calculated as ln(failure rate)$^{-1}$ (refs 39,40). The paired-pulse ratio was calculated as the ratio of peak amplitudes between the second and first EPSCs in a train. Effects of drug application were calculated after a perfusion time of at least 10 min. Statistical analysis was performed on cells pooled across mice.

Single unit recordings. *Data acquisition.* Neuropixels silicon probes (phase3, option1, 384 channels[41]) were used to record extracellular spikes from dmSC neurons in three male adult C57BL/6J wild-type mice. A craniotomy was made over the SC and sealed with Kwik-Cast, followed by attachment of a metal custom-made headplate and ground pin to the skull, using dental cement. At least 36 h after surgery, mice were placed on a plastic wheel and head-fixed at an angle of 30° from the anterior-posterior axis, parallel to an LCD monitor (Dell, 60-Hz refresh rate) centred 30 cm above the head. Before recording, the probe was coated with DiI (1 mM in ethanol, Invitrogen) for track identification and a wire was connected to the ground pin for external reference and ground. For recording, the probe was slowly inserted into the SC (AP: −0.5 to −0.7, ML: 0.4 to 0.8) to a depth of 2.8–3.0 mm and left in place for at least 20 min before the beginning of the recording session. Data was acquired using spikeGLX (https://github.com/billkarsh/SpikeGLX, Janelia Research Campus), high-pass filtered (300 Hz), amplified (500×), and sampled at 30 kHz. Sensory stimuli were delivered and synchronized using custom-made LabVIEW software. Mantis software (mantis64.com) and a PCIe-6353 board (National Instruments). Visual and auditory stimuli (98% contrast; 50% contrast; sound) were presented interleaved with a 1-min interval and a total of 30 presentations each. *Data analysis.* Analysis was performed in MATLAB 2017a. Raw voltage traces were band-pass filtered (300–5,000 Hz), spikes were detected and sorted automatically using JRCLUST[42], followed by manual curation. Only units with a clear absolute refractory period in the auto-correlogram were classified as single units. Firing-rate histograms were calculated as the average firing rate in bins of 1 ms for 30 consecutive trials, and subsequently smoothed. Units were considered to respond to the threat stimulus if their firing rate increased by at least 1 Hz in a 500-ms time-window from stimulus onset when compared to the baseline (500 ms before stimulus onset). Peak firing rates for each stimulus were calculated as the mean of a 30-ms time window centred on the time of the average peak firing rate of all responding units. Responses to 50% contrast visual stimuli were calculated on all units that responded to 98% contrast. For units showing persistent activity after stimulus offset, the time constant to decay to baseline was obtained by fitting a single exponential to the average firing rate histogram. Statistical analysis was performed on single units pooled from all mice.

Retrograde tracing. For monosynaptic rabies tracing[43,44] from the dPAG, TVA and RG were injected unilaterally into the dPAG[45] with an angled approach from the contralateral hemisphere to avoid infection of the SC in the target hemisphere (20°, AP: −0.45 to −0.5, ML: −0.6, DV: −2.2). EnvA-dG-RV-mCherry was injected into the dPAG vertically (AP: −0.4, ML: +0.5, DV: −2.1) 10–14 days later. Mice were perfused seven days post-rabies virus injection. Brains were cut at 100-μm thickness on a microtome (HM650V, Microm). All sections containing the PAG and SC were mounted in SlowFade Gold, and imaged using a confocal microscope (SP8, Leica). Tile scans of the entire section were acquired with a 25× water objective (Olympus) at five depths (10 μm apart) and maximum projections of these stacks were used for subsequent analysis. Cell counting was done manually (Cell counter plug-in, Fiji) in reference to the Franklin and Paxinos atlas[46]. To quantify the position of presynaptic SC cells along the mediolateral axis, the coordinates of the counted cells were normalized to the medial and lateral extents of the SC for each brain slice, and a kernel density estimation was performed (Scikit-learn, Python). For retrograde tracing from the dmSC, rAAV2-retro-mCherry was injected unilaterally. AAV2-CamkII-GFP was co-injected to label the injection site in two out of three brains. Mice were euthanized 14–18 days afterwards and their brains processed as described above. Additionally, rabies tracing from the mSC was performed in three mice, and as described above. Every third section along the rostrocaudal axis of the SC was imaged with on an Axio Imager 2 (Zeiss) and presynaptic cells in the dPAG and auditory cortex were counted manually.

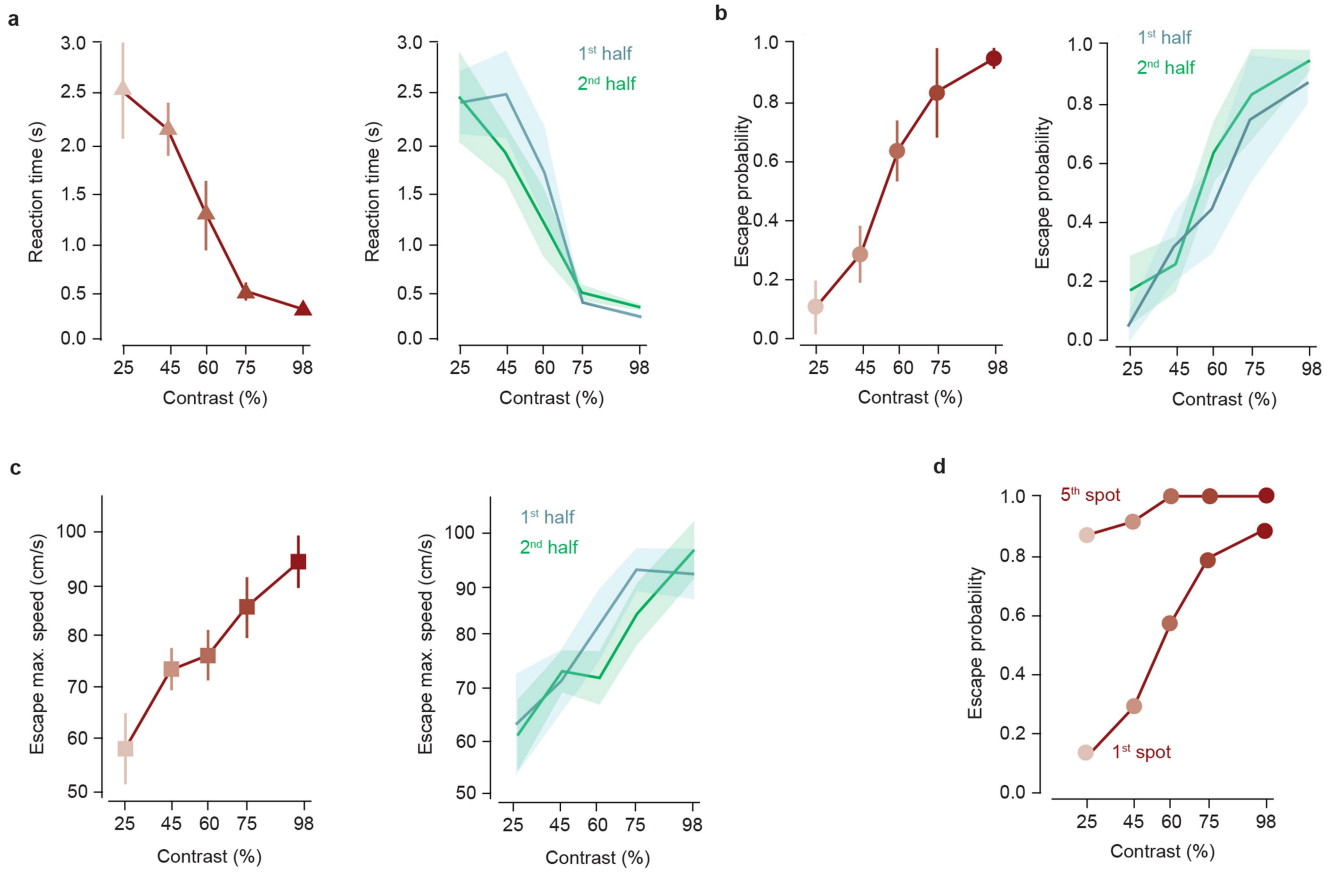
Histological quantifications. To estimate the fraction of VGluT2$^+$ cells in a target area that were infected with viral vectors, we compared the density of infected cells in VGluT2-Cre mice at the implant site, to control densities quantified using the VGluT2::eYFP reporter line. Optogenetic vectors infected 86 ± 6% for dPAG and 95 ± 9% for mSC; GCaMP6s infected 90 ± 8% for dPAG and 86 ± 1% for mSC; hM4D infected 93 ± 15% for mSC. The placement of optic fibres, GRIN lenses and cannulae was assessed histologically based on their tract and tip location, and their tip locations are illustrated in the respective sections of the mouse brain atlas[46] (see Extended Data Figs. 2, 4, 6 and 10).

General data analysis. Data analysis was performed using custom-written routines in Python 2.7 and custom code will be made available on request. Data are reported as mean ± s.e.m. unless otherwise indicated. Statistical comparisons using the significance tests stated in the main text were made in SciPy Stats and GraphPad Prism, and statistical significance was considered when $P < 0.05$. Data were tested for normality with the Shapiro–Wilk test, and a parametric test used if the data were normally distributed, and a non-parametric otherwise, as detailed in the text next to each comparison.

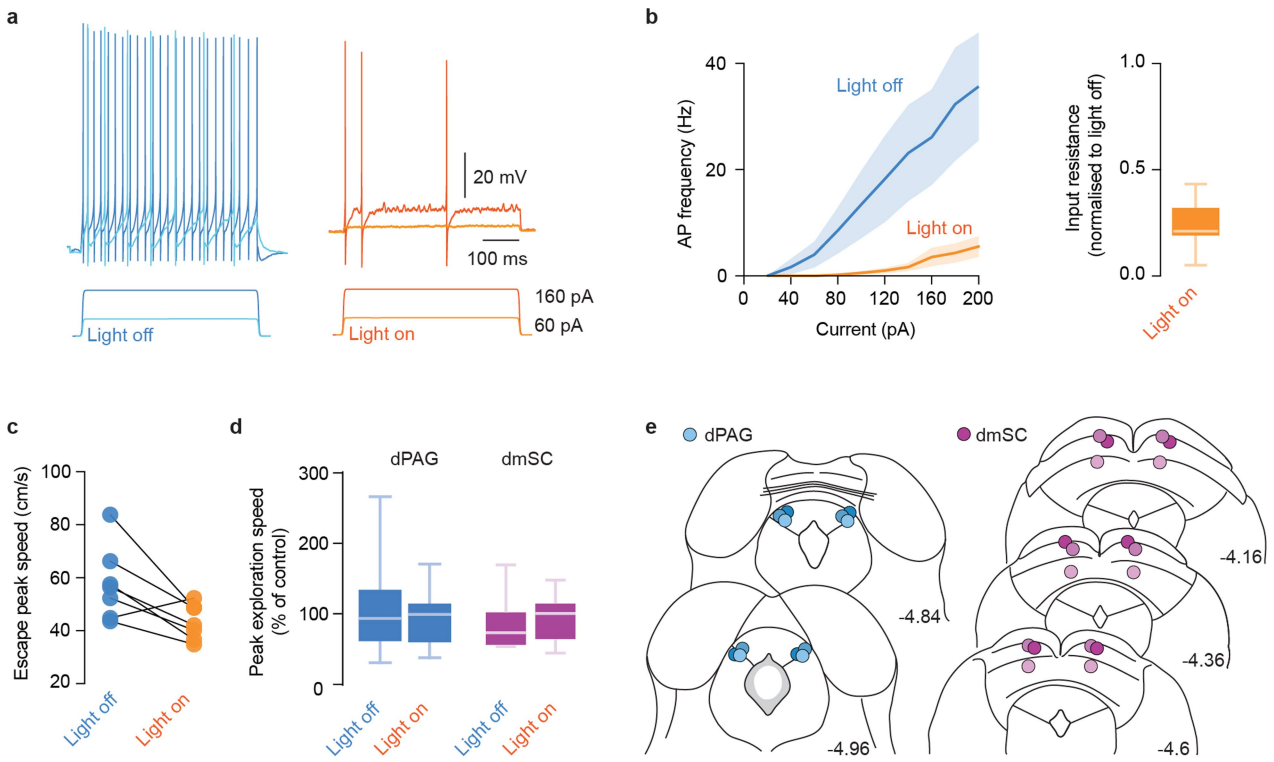
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The datasets generated and/or analysed in this study are available from the corresponding author upon reasonable request.

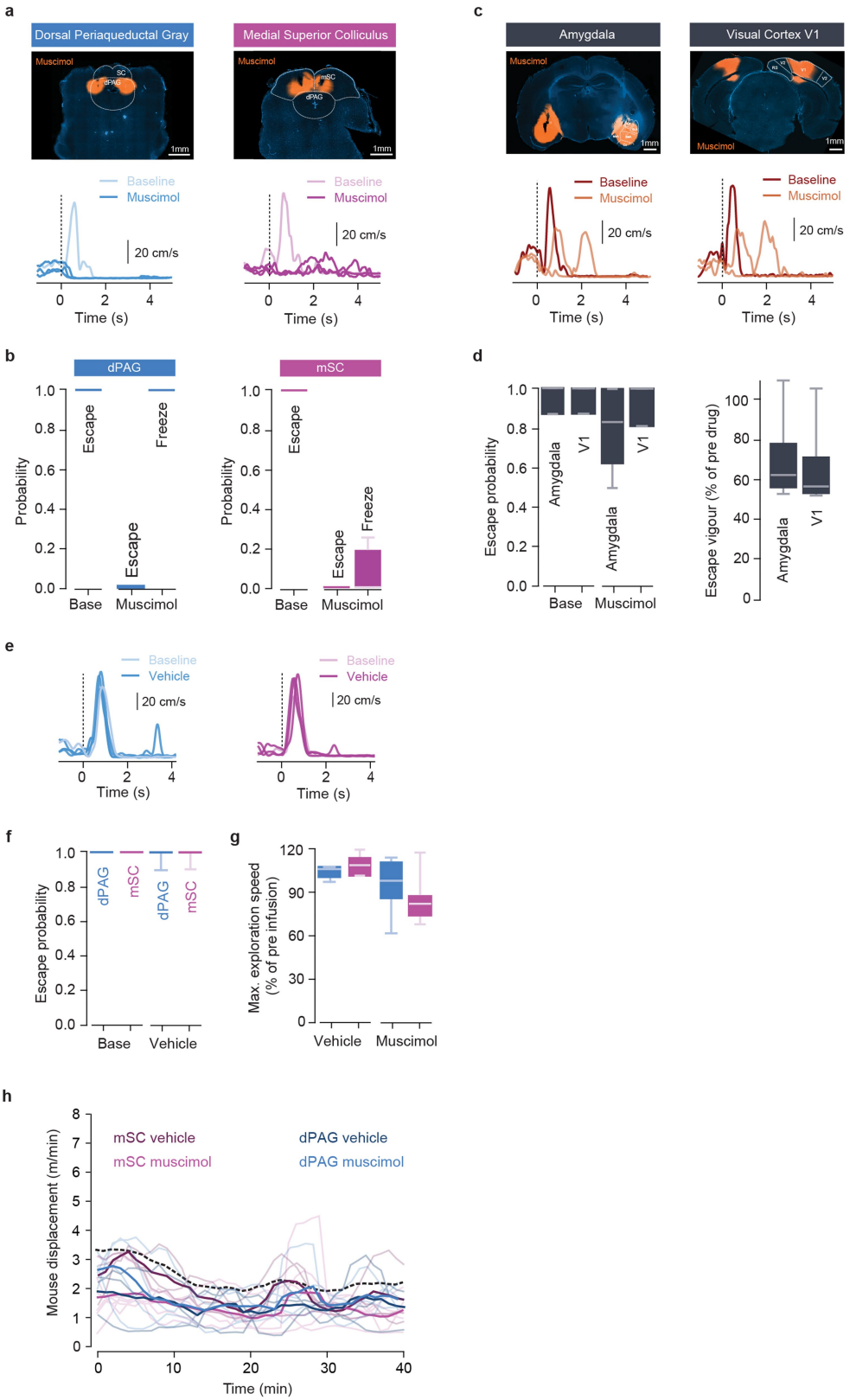
29. Vong, L. et al. Leptin action on GABAergic neurons prevents obesity and reduces inhibitory tone to POMC neurons. *Neuron* **71**, 142–154 (2011).
30. Vélez-Fort, M. et al. The stimulus selectivity and connectivity of layer six principal cells reveals cortical microcircuits underlying visual processing. *Neuron* **83**, 1431–1443 (2014); erratum **84**, 238 (2014).
31. Vélez-Fort, M. et al. A circuit for integration of head- and visual-motion signals in layer 6 of mouse primary visual cortex. *Neuron* **98**, 179–191.e6 (2018).
32. Tervo, D. G. R. et al. A designer AAV variant permits efficient retrograde access to projection neurons. *Neuron* **92**, 372–382 (2016).
33. Mongeau, R., Miller, G. A., Chiang, E. & Anderson, D. J. Neural correlates of competing fear behaviors evoked by an innately aversive stimulus. *J. Neurosci.* **23**, 3855–3868 (2003).
34. Resulaj, A., Kiani, R., Wolpert, D. M. & Shadlen, M. N. Changes of mind in decision-making. *Nature* **461**, 263–266 (2009).
35. Ghosh, K. K. et al. Miniaturized integration of a fluorescence microscope. *Nat. Methods* **8**, 871–878 (2011).
36. Guizar-Sicairos, M., Thurman, S. T. & Fienup, J. R. Efficient subpixel image registration algorithms. *Opt. Lett.* **33**, 156–158 (2008).
37. Aravanis, A. M. et al. An optical neural interface: *in vivo* control of rodent motor cortex with integrated fiberoptic and optogenetic technology. *J. Neural Eng.* **4**, S143–S156 (2007).
38. Huber, D. et al. Sparse optical microstimulation in barrel cortex drives learned behaviour in freely moving mice. *Nature* **451**, 61–64 (2008).
39. Isaacson, J. S. & Walmsley, B. Counting quanta: direct measurements of transmitter release at a central synapse. *Neuron* **15**, 875–884 (1995).
40. del Castillo, J. & Katz, B. Quantal components of the end-plate potential. *J. Physiol. (Lond.)* **124**, 560–573 (1954).
41. Jun, J. J. et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017).
42. Jun, J. J. et al. Real-time spike sorting platform for high-density extracellular probes with ground-truth validation and drift correction. Preprint at https://www.bioRxiv.org/content/early/2017/01/30/101030 (2017).
43. Wall, N. R., Wickersham, I. R., Cetin, A., De La Parra, M. & Callaway, E. M. Monosynaptic circuit tracing *in vivo* through Cre-dependent targeting and complementation of modified rabies virus. *Proc. Natl Acad. Sci. USA* **107**, 21848–21853 (2010).
44. Wickersham, I. R. et al. Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. *Neuron* **53**, 639–647 (2007).
45. Franklin, T. B. et al. Prefrontal cortical control of a brainstem social behavior circuit. *Nat. Neurosci.* **20**, 260–270 (2017).
46. Franklin, K. B. J. & Paxinos, G. *The Mouse Brain in Stereotaxic Coordinates* 3rd edn (Academic Press, 2008).



Extended Data Fig. 1 | Behaviour metrics computed over single mice. **a**–**c**, Summary plots for escape behaviour metrics calculated for each mouse individually and averaged. Plots on the left were obtained with data from all trials, and in the plots on the right, trials for each contrast were split in half and the behaviour metrics calculated for each half. There is a significant dependency on contrast for all metrics (reaction time, **a**: $P = 3.5 \times 10^{-8}$; escape probability, **b**: $P = 2.1 \times 10^{-7}$; escape vigour, **c**: $P = 1.6 \times 10^{-6}$, repeated measures ANOVA), and no significant difference between the metrics calculated using the first and second half of the trials ($P > 0.4$ for a main effect of trial group in all comparisons, two-way repeated measures ANOVA), indicating that behavioural performance was stable across repeated presentations of the stimulus. Error bars and shaded areas are s.e.m. **d**, Escape probability after the first (as shown in Fig. 1e, calculated by pooling all data) and fifth spot, during the presentation of five consecutive expanding spots.

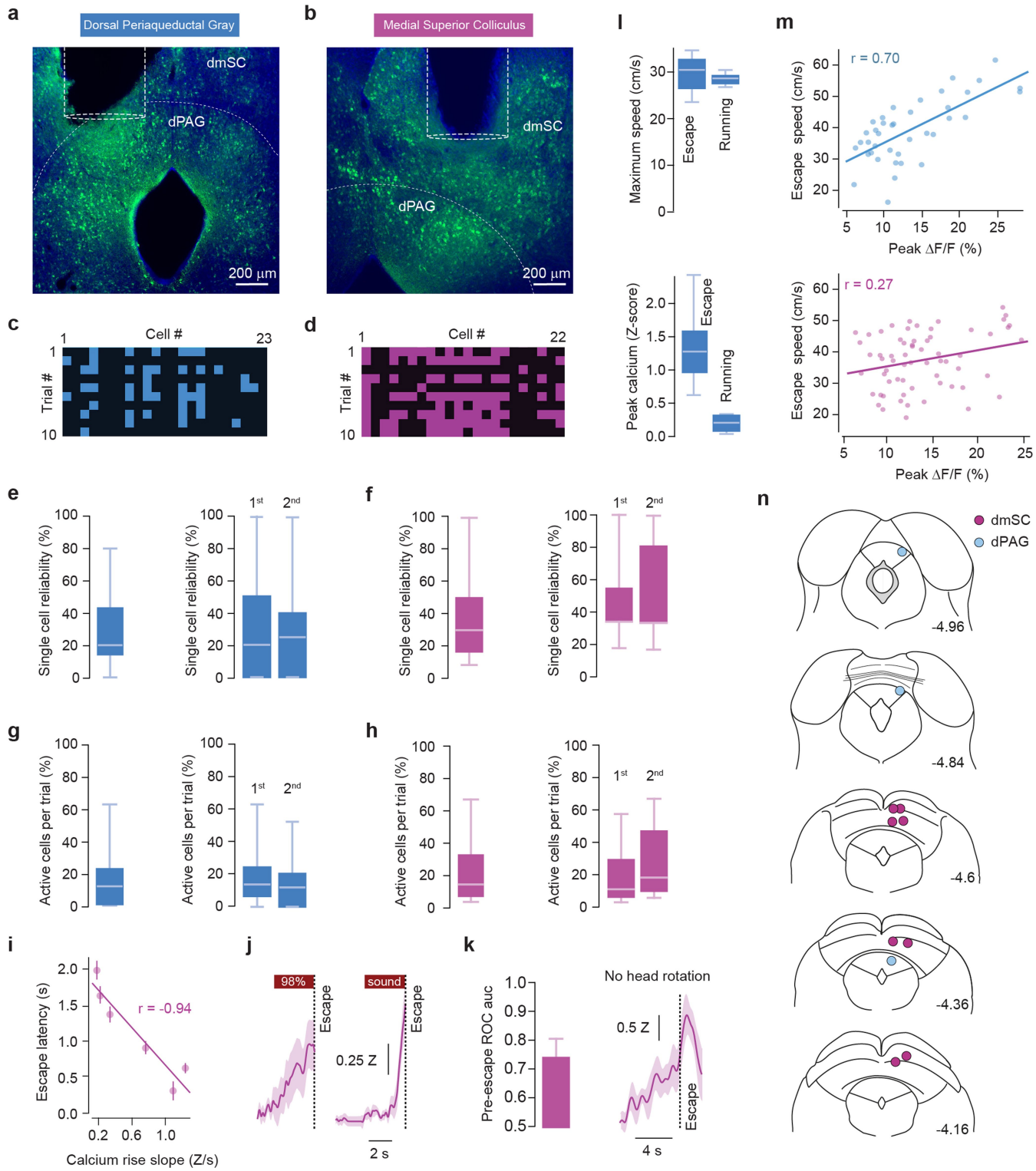


Extended Data Fig. 2 | iChloC activation strongly reduces neuronal firing and disrupts defensive behaviour without affecting basal locomotion. a, Example voltage traces showing a VGluT2^{+} dmSC neuron expressing iChloC responding to current steps in control conditions (light off, left) and during continuous illumination with 473-nm light (light on, right). **b**, Summary of the relationship between current injection and action potential firing showing a strong reduction in firing upon illumination (left, average 87.9 ± 3% reduction across all steps, $P = 1.7 \times 10^{-9}$ for a main effect of light, two-way repeated measures ANOVA; $P < 0.05$ for simple effects of light on current steps larger than 100 pA), as well as a strong reduction in input resistance (right, 73.2 ± 3% reduction, $P = 1.23 \times 10^{-8}$, *t*-test). Summary data are pooled from 6 dPAG and 3 dmSC cells. c, For the 18% of trials in which VGluT2^{+} mice expressing iChloC in the dmSC escape from threat stimuli during continuous illumination (light on), the vigour of escape is significantly lower (77 ± 7% of light off) when compared to escapes elicited without iChloC activation (light off; $n = 7$ trials, $n = 6$ out of 9 mice, $P = 0.0253$, paired *t*-test). d, Movement during exploration is not affected by iChloC activation in dPAG- or dmSC-targeted mice in the absence of threat, quantified as the maximum speed in the 5-s stimulation period (light on) or control period (light off) as a percentage of the 5-s pre-stimulation period ($P = 0.8767$ for dPAG, $P = 0.3443$ for dmSC, *U*-test). e, Optic fibre placements for all experiments in dPAG ($n = 6$ mice, blue circles) and dmSC ($n = 9$ mice, magenta circles), coordinates are in mm and from bregma. Mouse brain images adapted from ref. [46] and reproduced with permission from Elsevier.



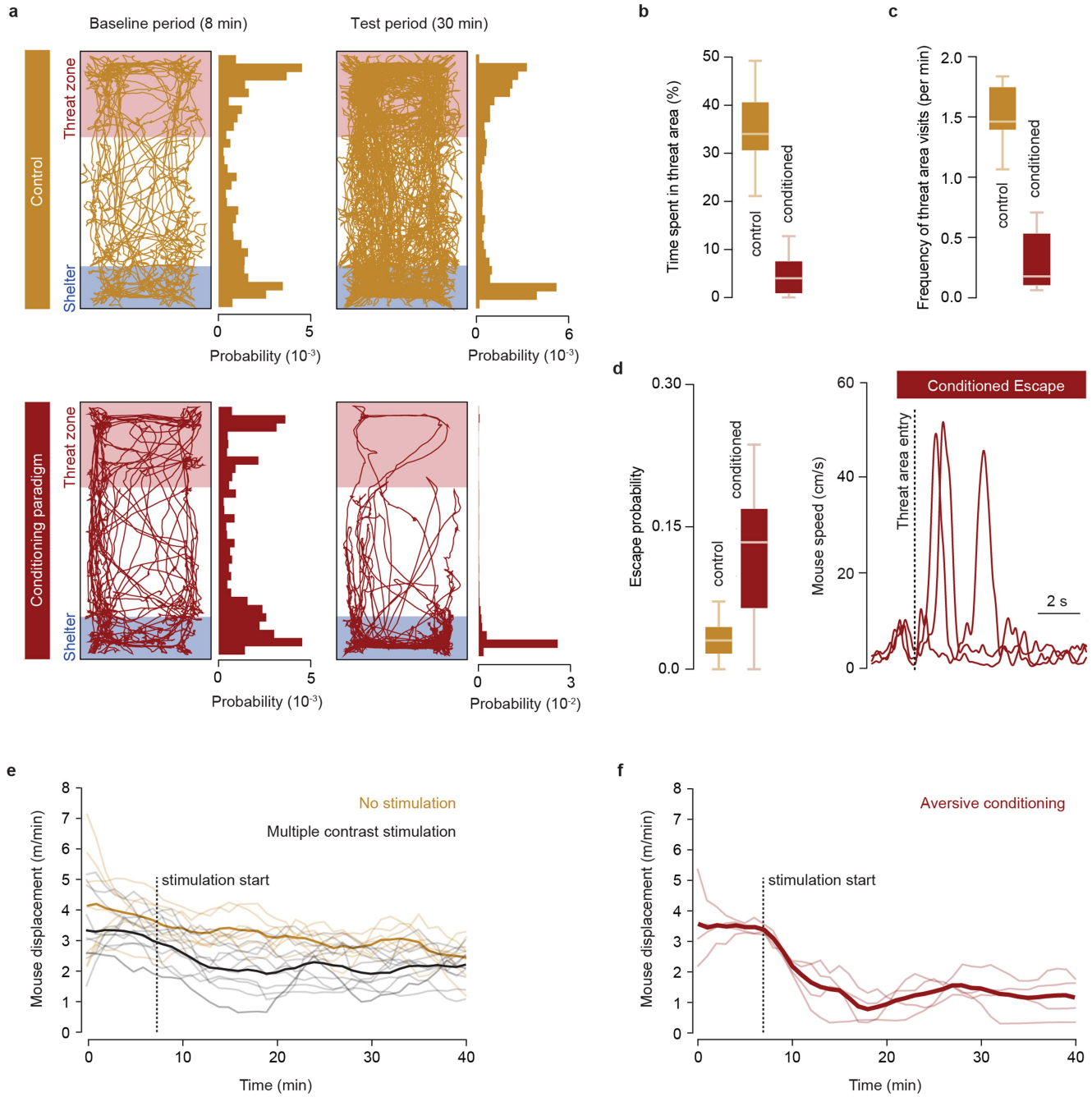
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Muscimol inactivation of dPAG and mSC abolishes escape while V1 and amygdala have a modulatory effect on escape behaviour. a, Top, example images of muscimol infusion in the dPAG (left) and mSC (right), and respective speed traces in response to a threatening visual stimulus (bottom) showing a switch from escape to freezing after dPAG inactivation and a loss of defensive responses after mSC muscimol inactivation. **b**, Summary quantification of the effect of muscimol infusion on threat-evoked defensive behaviour probability in the dPAG (left; $n = 7$ mice, $P = 0.0001$ for escape and $P = 0.00025$ for freezing, *U*-tests) and mSC (right; $n = 10$ mice, $P = 0.00021$ for escape and $P = 0.051$ for freezing, *U*-tests). **c**, Top, images of bilateral muscimol infusion in the amygdala (left) and visual cortex area V1 (right). Respective speed traces during threatening visual stimulus presentation (bottom) show that mice still engage in escape behaviour, but with reduced vigour. **d**, Summary quantification for escape probability (left) and vigour (right) after amygdala and V1 acute inactivation (amygdala: $n = 4$ mice, $P = 0.37$ for escape probability, *U*-test; $P = 0.01$ for escape vigour, two-tailed *t*-test; V1: $n = 4$ mice, $P = 0.5$ for escape probability, *U*-test; $P = 0.01$ for escape vigour, two-tailed *t*-test). **e**, Example speed traces showing that vehicle infusion in the mSC and dPAG does not change threat-evoked escape probability, and respective summary quantification. **f**, Infusion of mSC and dPAG with vehicle does not affect escape probability (mSC; $n = 5$ mice, $P = 0.21$, *U*-test; dPAG; $n = 5$ mice, $P = 0.21$, *U*-test). **g**, Infusion of mSC and dPAG with muscimol or vehicle does not affect running speed during exploratory behaviour (mSC: $P = 0.8$ for vehicle, $P = 0.22$ for muscimol; dPAG: $P = 0.28$ for vehicle, $P = 0.75$ for muscimol, paired *t*-tests). **h**, Profile of exploratory behaviour for behavioural sessions lasting at least 40 min, after injection of vehicle or muscimol in the mSC and dPAG. The displacement over time for all conditions is not significantly different to the profile for multiple trials of visual threat stimulation in control conditions (dashed black line, same data as shown in Extended Data Fig. 5e; $P > 0.1$ for all comparisons with control, two-tailed *t*-test). Thin lines show individual mice and thick lines show the dataset mean. Box-and-whisker plots show median, IQR and range.

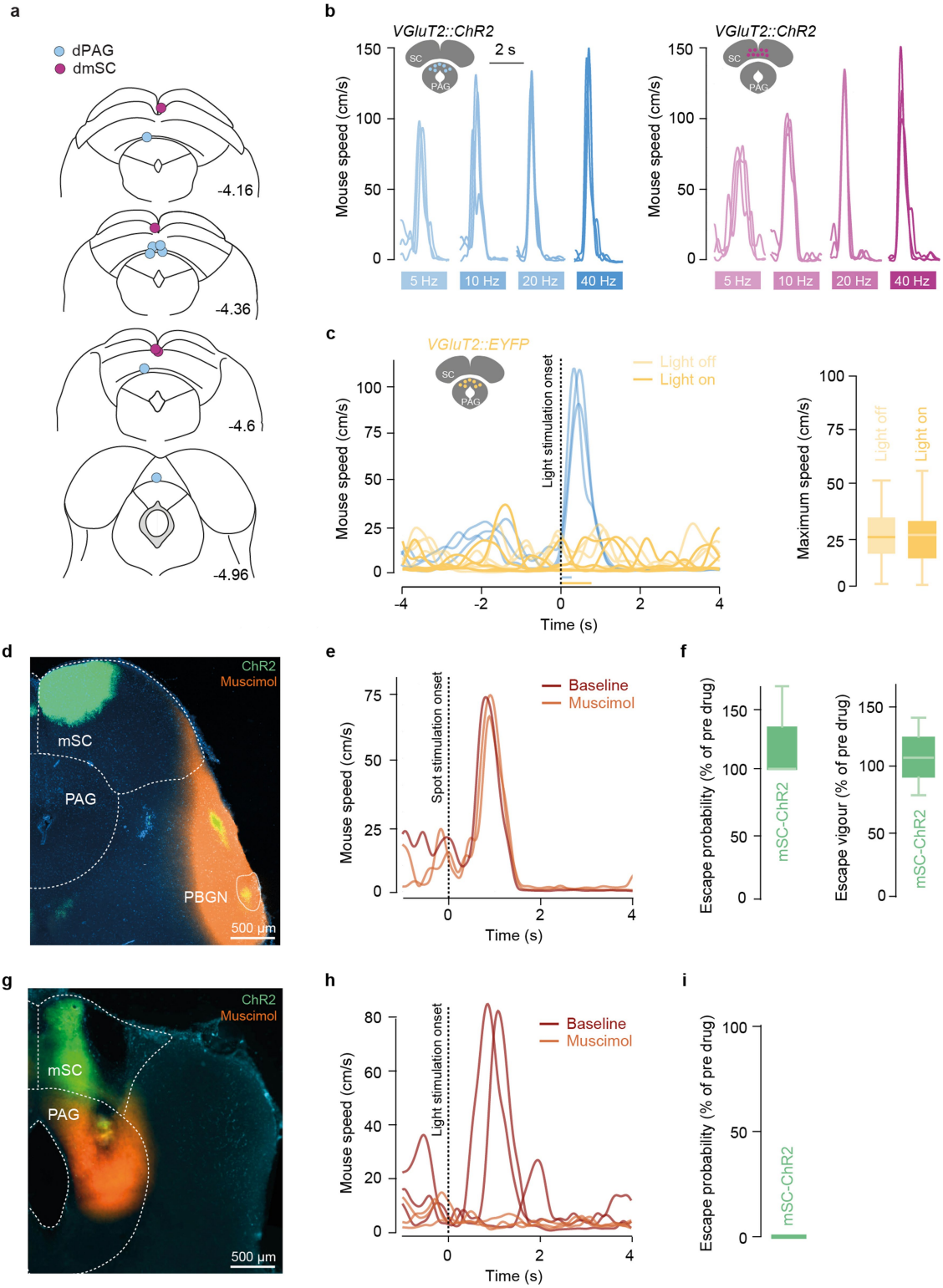


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | The reliability and fraction of active cells is stable over multiple trials of calcium imaging, activity in the dmSC does not reflect head rotation and rises with different slopes, and dPAG activity is specific to escape. a, **b**, Example images of GCaMP6s expression in VGluT2^{+} cells (green), with schematic showing GRIN lens placement in the dPAG (**a**) and dmSC (**b**). **c**, **d**, Raster plots showing active (colour squares) and non-active cells (black squares) in a single FOV imaged over multiple trials. A total of 8 FOVs were imaged in the dPAG (**c**) with a mean of 18 cells per FOV (range = 7–30) and 11 trials per FOV; and in the dmSC (**d**), 11 FOVs were imaged with a mean of 20 cells per FOV (range = 7–31) and 20 trials per FOV. There was a mean of 7 escape-responding cells per dPAG FOV and 16 escape-responding cells per dmSC FOV. **e**, **f**, Reliability of escape-responding cells showing a response over multiple trials for all trials (left) and for the first and second half of trials separately (right). Mean reliability across all trials was $28 \pm 3\%$ for dPAG and $35 \pm 3\%$ for dmSC, and stable over multiple trials ($P = 0.44$ for dPAG, $P = 0.11$ for dmSC, comparison between the two groups of trials, *U*-test). **g**, **h**, Fraction of all cells in a FOV that were active on each trial for all trials (left) and for the first and second half of trials separately (right). The active fraction across all trials was $14 \pm 3\%$ for dPAG and $23 \pm 6\%$ for dmSC, and stable over multiple trials ($P = 0.21$ for dPAG, $P = 0.08$ for dmSC, comparison between the two groups of trials, *U*-test). **i**, Correlation between the rise slope of the population activity and escape latency ($n = 75$ trials, $P = 0.0048$, Pearson's *r*). **j**, Average population calcium signal in the dmSC for escape trials in response to 98% contrast spots and sound stimuli. The slope of the signal rise is steeper for sound-evoked escape. **k**, Left, ROC AUC for the dmSC signal before spontaneous escape onset after conditioning (AUC at escape onset = 0.74, significantly above chance 2.1 s before escape, $n = 57$ trials). Right, average population calcium signal in the dmSC during threat-evoked escape trials where the mouse was already facing the shelter and therefore did not rotate the head ($n = 5$ trials). **l**, Summary quantification of dPAG population calcium signals during threat-evoked escape and spontaneous foraging running bouts of similar speed (top; $n = 6$ escape trials and $n = 6$ running bouts, speed not significantly different, $P = 0.64$, *t*-test), showing that activity increase in the dPAG is specific for escape (bottom; $P = 0.0018$, *t*-test). Shaded areas show s.e.m., box-and-whisker plots show median, IQR and range. **m**, Correlation between the population activity of dPAG (top; $n = 39$ trials, $P = 6.7 \times 10^{-7}$, Pearson's *r*) and dmSC (bottom; $n = 64$ trials, $P = 0.04$, Pearson's *r*) and escape speed. Each data point is a single trial. **n**, Placement of GRIN lenses in the dmSC (magenta circles) and dPAG (blue circles), coordinates are in mm and from bregma. Mouse brain images adapted from ref. [46] and reproduced with permission from Elsevier. Box-and-whisker plots show median, IQR and range.

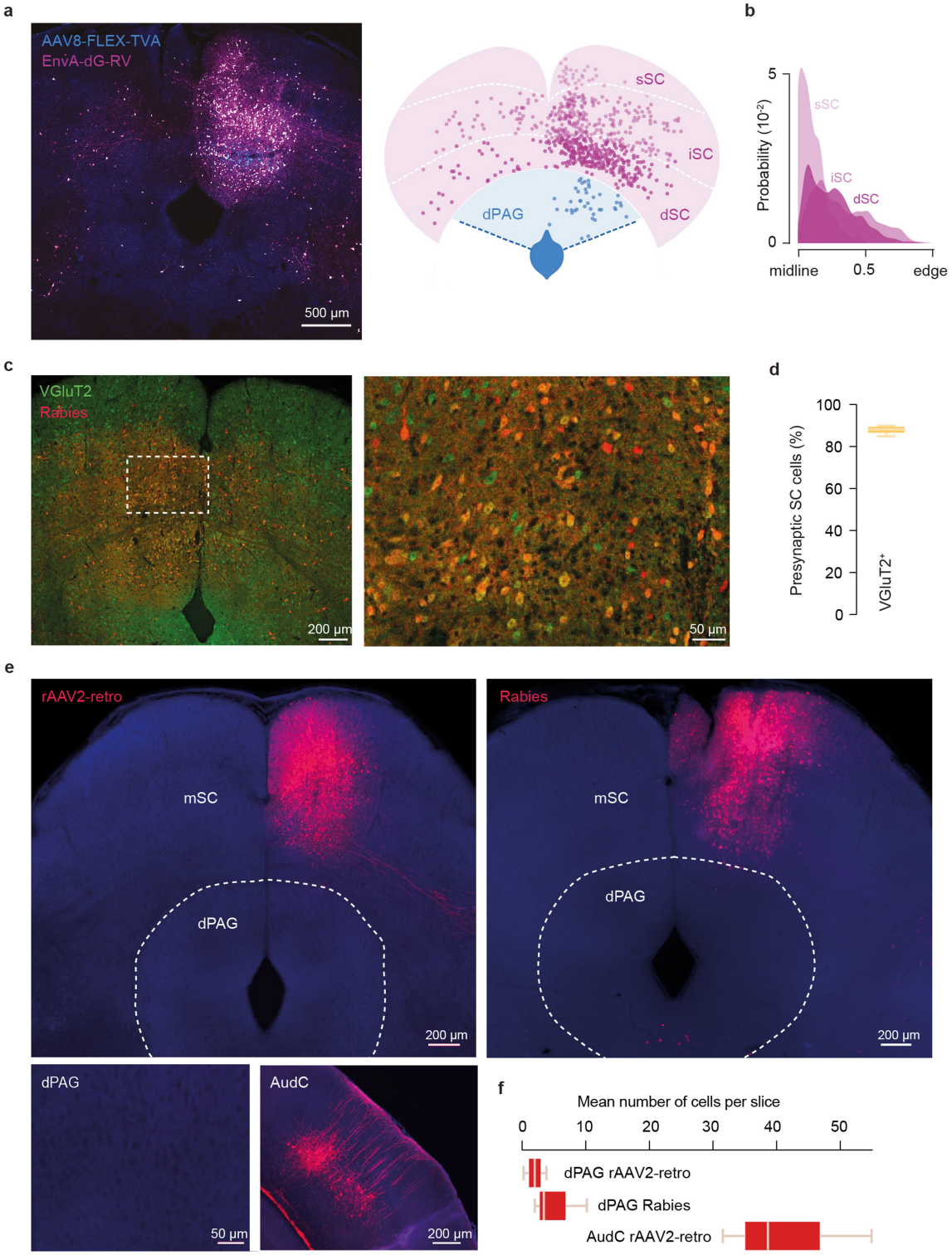


Extended Data Fig. 5 | Repeated high-contrast visual stimulation causes place aversion, reduction in exploration and spontaneous escape. **a**, Traces and probability distributions for the location of two example mice during free exploration (top), and before and after a high-contrast visual stimulation conditioning paradigm (bottom), showing avoidance of the threat area after conditioning (bottom right). **b**, Time spent in the threat area decreases with aversive conditioning ($35.1 \pm 3.5\%$ for naive mice versus $5.1 \pm 2.0\%$ after conditioning, $n = 7$ mice, $P = 2.2 \times 10^{-5}$, two-tailed t-test). **c**, The frequency of visits to the threat area by the mice decreases significantly after conditioning (1.51 ± 0.10 visits per min for naive mice versus 0.30 ± 0.12 after conditioning, $n = 7$ mice, $P = 1 \times 10^{-4}$, two-tailed t-test). **d**, Summary quantification of spontaneous escape probability (left) and single trial speed traces from three mice (right) showing spontaneous escape after conditioning ($P_{\text{spontaneous escape}}$ $3.2 \pm 0.8\%$ for naive mice, $n = 7$ mice, and $12.2 \pm 2\%$ after conditioning, $n = 13$ mice; $P = 0.004$, two-tailed t-test). **e**, Profile of exploratory behaviour during behavioural sessions of multiple contrast stimulation (black, data taken from the mice that generated the dataset for Fig. 1) with no stimulation for comparison (orange). Exploration decays over time and the decay is accelerated by visual stimulation, but the two curves are not significantly different over time (2.4 ± 0.3 m min^{-1} at 40 min for control versus 2.0 ± 0.3 with visual stimulation, $P = 0.16$, two-tailed t-test). **f**, Same quantification as in **e** for sessions of aversive conditioning. Aversive conditioning significantly reduces exploratory behaviour (1.2 ± 0.3 m min^{-1} after conditioning, $P = 0.018$ versus no stimulation and $P = 0.039$ versus multiple contrast stimulation, two-tailed t-test). Thin lines show individual mice monitored for 40 min and thick lines show the dataset mean. Box-and-whisker plots show median, IQR and range.



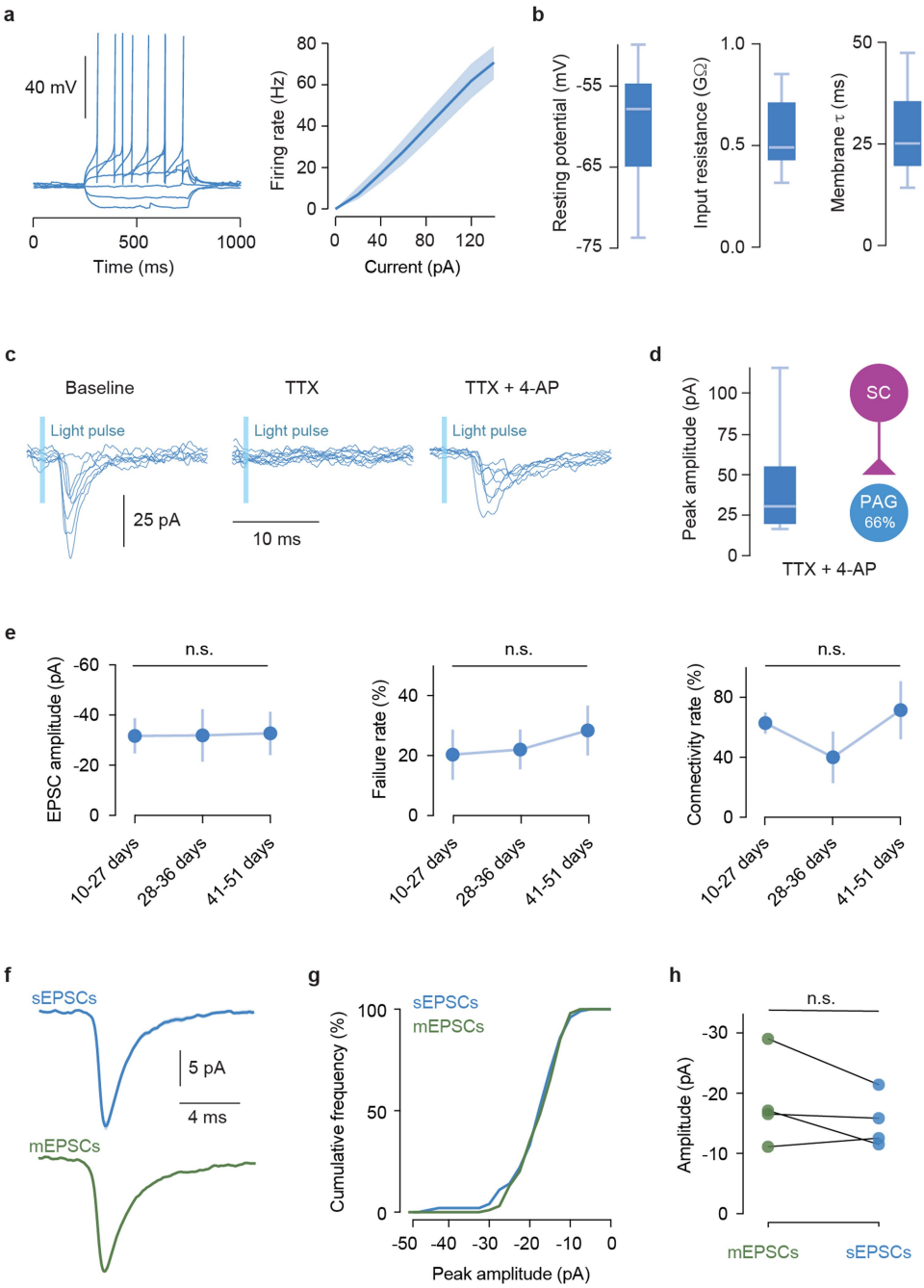
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Optogenetic activation of dPAG and mSC elicits escape over a range of frequencies, and mSC VGluT2::ChR2-evoked escape is abolished by inactivating the PAG, but not the PBGN. **a**, Optic-fibre placements for ChR2 stimulation in the dmSC (magenta circles) and dPAG (blue circles), coordinates are in mm and from bregma. Mouse brain images adapted from Franklin and Paxinos[46] and reproduced with permission from Elsevier. **b**, Example speed traces for dPAG (left) and mSC (right) ChR2 stimulation at different frequencies (10 pulses) and high light intensities, showing robust escape behaviour for 5 to 40 Hz stimulation. **c**, Left, speed traces for 473-nm light stimulation (40 Hz, 30 pulses) of one mouse expressing eYFP in the dPAG (dark green), showing no change in running speed. Light green traces show similar speed profiles for the same mouse entering the stimulation area with the light off. Blue dashed traces are from a different mouse expressing ChR2 in the dPAG (40 Hz, 10 pulses), for comparison. Right, summary data for eYFP control stimulation in dPAG (running speed not significantly different between laser on and off, $n = 236$ trials from 3 mice, $P = 0.48$, *U*-test). **d**, Image showing expression of ChR2–eYFP in the mSC (green) with projections to the PBGN (yellow) and muscimol infusion (orange). **e**, Speed traces for spot-evoked escape responses from one mouse before and after acute PBGN inactivation. **f**, Summary data for escape probability and vigour during mSC optogenetic stimulation and PBGN acute inactivation, showing no difference ($n = 3$ mice, $P = 0.80$ for escape probability; $P = 0.70$ for escape vigour, *U*-test). **g**, Image showing expression of ChR2–eYFP in the mSC (green) and muscimol infusion in the PAG (orange). **h**, **i**, Speed traces (**h**) and summary data (**i**) showing that mSC ChR2-evoked escape is abolished by PAG acute inactivation ($n = 3$ mice, $P = 0.0297$ for probability, *U*-test). Box-and-whisker plots show median, IQR and range.

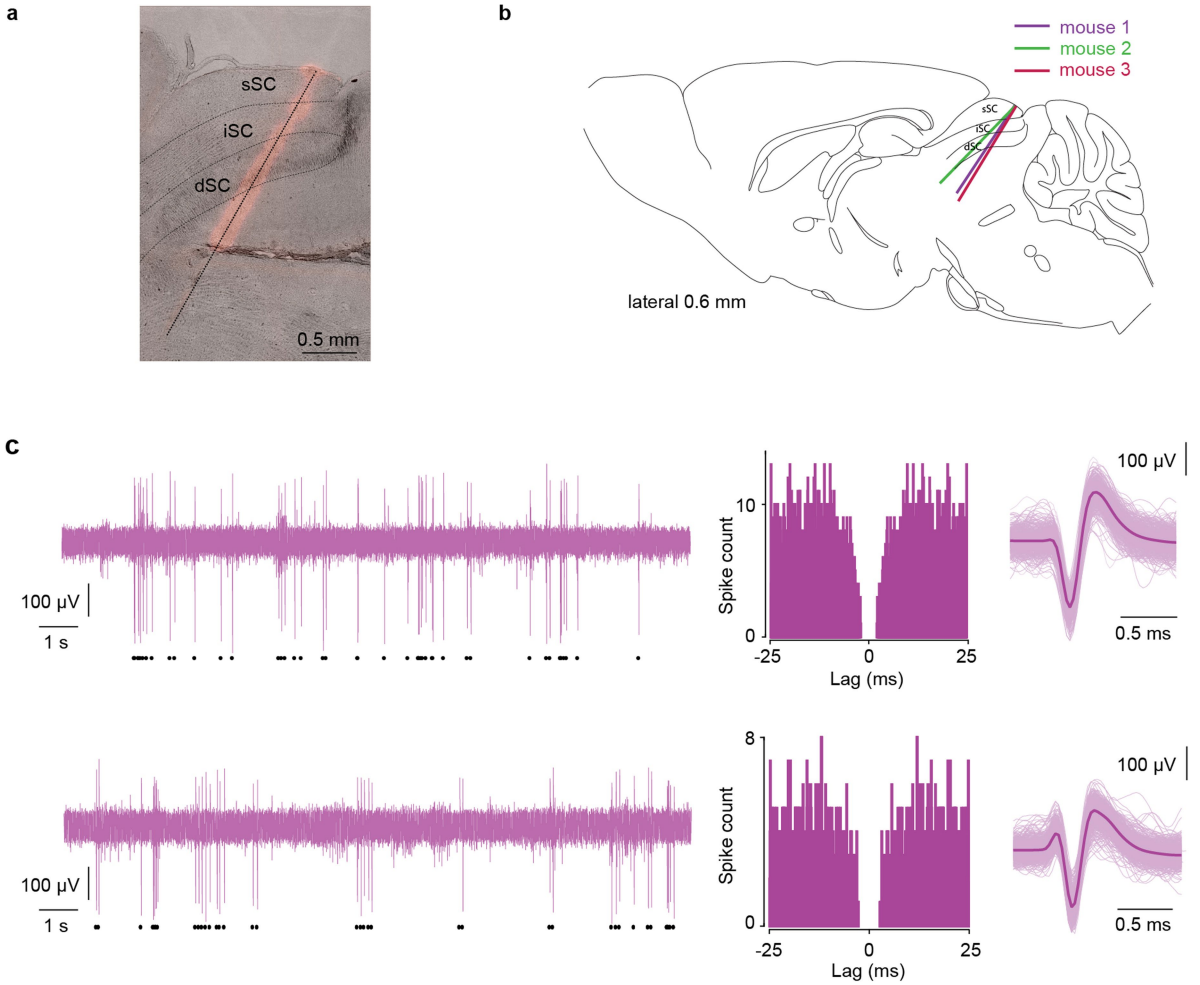


Extended Data Fig. 7 | See next page for caption.

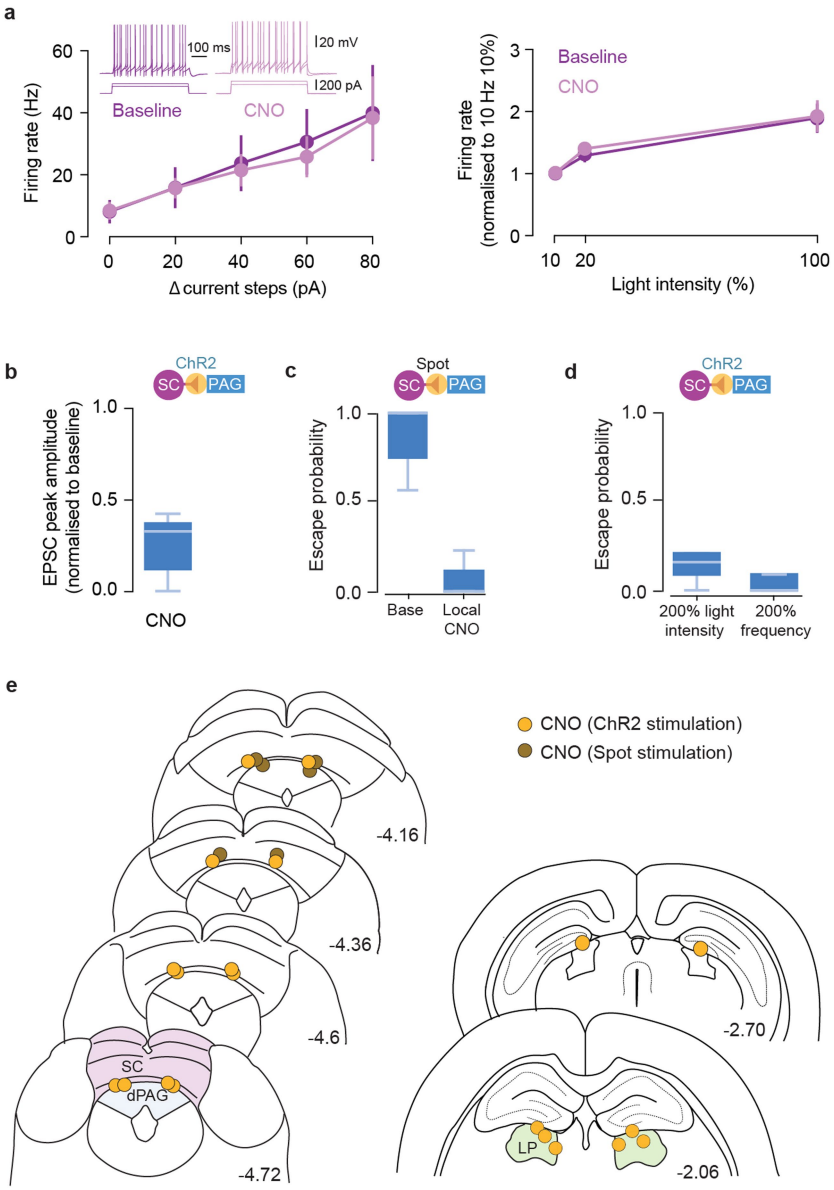
Extended Data Fig. 7 | dPAG neurons receive input from mainly excitatory cells in the SC and do not project back to the SC. a, Image showing starter dPAG VGlut2^{+} cells expressing both TVA–GFP and RV–mCherry and presynaptic cells expressing RV–mCherry only (left), and corresponding schematic (right) illustrating the position of starter dPAG (blue) and presynaptic SC cells (pink) across deep, intermediate and superficial SC layers (same as shown in Fig. 4a). **b**, Kernel density estimation curves for the axial position of presynaptic SC cells for each layer ($82.9 \pm 2.6\%$ of 1,770 cells are located within the medial bisection of ipsilateral SC, $n = 3$ mice). **c**, Image showing presynaptic cells in the mSC infected with rabies virus (red) from starter neurons in the dPAG of a VGluT2::eYFP mouse (left). Box indicates area magnified shown on the right. Yellow cells are VGluT2^{+} mSC presynaptic neurons. **d**, Summary quantification of the percentage of presynaptic cells in the mSC that express VGluT2^{+} (mean $= 87.9 \pm 1.0\%$, $n = 4$ mice). **e**, Image showing injection of rAAV2-retro in the mSC (left) and no retrogradely labelled cells in the dPAG (bottom, left), while retrograde labelling is present in the auditory cortex for comparison (bottom, right). Similarly, rabies virus injected in the mSC shows a lack of presynaptic cells in the dPAG (right), suggesting a predominantly feed-forward connectivity arrangement between the mSC and dPAG (note, however, that it cannot be excluded that both rAAV2-retro and rabies display selective tropism that prevents labelling of dPAG neurons). **f**, Summary quantification for retrogradely labelled cells in the dPAG and auditory cortex after mSC rAAV2-retro ($n = 3$ mice) or rabies infection ($n = 3$ mice). Box-and-whisker plots show median, IQR and range.



Extended Data Fig. 8 | Biophysical properties of excitatory dPAG neurons and synaptic properties of the dmSC-dPAG excitatory connection. a, Example trace of current step injections in a VGluT2$^+$ dPAG cell (left) and summary current–frequency relationship (right, shaded area is s.e.m.). **b**, Summary quantification of resting membrane potential (mean = −61.4 ± 2.15), input resistance (mean = 0.55 ± 0.05 GΩ) and membrane time constant (mean = 28.3 ± 3 ms) for VGluT2$^+$ dPAG cells (n = 14 cells, n = 7 mice). **c**, Example current traces for one dPAG VGluT2$^+$ cell showing optogenetically evoked EPSCs from the dmSC (left) that are blocked by TTX (middle) and recovered by 4-AP (right), confirming the presence of a monosynaptic connection. **d**, Summary data for peak dmSC–dPAG EPSC amplitudes and connectivity rate in the presence of TTX and 4-AP. **e**, Summary data showing that the properties of the dmSC–dPAG connection do not change with number of days after viral transfection of ChR2, and remain weak and unreliable (n = 15 mice, P = 0.78, 0.51 and 0.33 for amplitude, failure rate and connectivity rate, respectively, Kruskal–Wallis test). Plots show mean and s.e.m. **f**, Average waveforms for sEPSCs and mESPCs (recorded in TTX) in one cell, and respective cumulative histogram for peak amplitudes. **g**, Peak amplitude of sEPSCs and mEPSCs is not significantly different (n = 4 cells, P = 0.18, 0.79, 0.9 and 0.36 respectively, Kolmogorov–Smirnov test for 100 events in each condition per cell). Box-and-whisker plots show median, IQR and range.



Extended Data Fig. 9 | Silicon probe anatomical placement and examples of dmSC single units. a, Example image showing the track left by one probe stained with DiI, superimposed on a bright-field image of a 30-µm sagittal slice. **b**, Schematic illustrating the probe track in each mouse (sagittal section, 0.6 mm lateral to the midline). Mouse brain image adapted from Franklin and Paxinos[46] and reproduced with permission from Elsevier. **c**, Two examples of dmSC single units (top and bottom). Left, raw voltage trace from the channel with the strongest signal for the unit of interest (black symbols below indicate all spikes detected for the unit). Middle, auto-correlogram of spike times calculated in bins of 1/30 ms. Right, superimposed action potential waveforms chosen randomly from the whole recording (light colour) and average waveform (dark colour).



Extended Data Fig. 10 | Controls and cannulae placements for chemogenetic inactivation experiments. a, Summary in vitro data for hM4D-neurexin/ChR2-expressing VGluT2^{+} dmSC neurons before (baseline) and after CNO application (CNO), showing no effect of CNO on action potential firing in response to current injection (left, $n = 6$ cells, $P = 0.8738$ for main effect of CNO, two-way repeated measured ANOVA; inset shows example traces to two current steps) or to 473-nm light-evoked ChR2 activation (right, $n = 9$ cells, $P = 0.7006$ for main effect of CNO, two-way repeated measured ANOVA). Error bars are s.e.m. **b**, Application of CNO reduces dmSC–dPAG excitatory synaptic transmission by $71 \pm 7\%$ ($n = 10$ cells, $P = 6.19 \times 10^{-6}$, two-tailed t-test between baseline and CNO). **c**, Disrupting mSC–dPAG synapses with CNO microinfusion in behaving mice blocks visually evoked escape behaviour ($n = 3$ mice, $P = 0.036$, U-test). **d**, Doubling the intensity or frequency of mSC stimulation while locally blocking mSC-dPAG synapses is not sufficient to rescue escape behaviour ($n = 5$ mice, $P = 0.11$ for intensity, U-test; $P = 0.42$ for frequency, U-test; both comparisons against escape probability after local block in baseline conditions shown in Fig. 4l). **e**, Cannula placements for local inactivation experiments with CNO at the SC–PAG synapse (left) and at the SC–LP synapse (right). The tip of the internal cannulae is indicated by yellow circles (for experiments with optogenetic stimulation of dmSC VGluT2^{+} cells) and brown circles (for experiments with visual stimulation). Coordinates are in mm and from bregma. Mouse brain images adapted from Franklin and Paxinos[46] and reproduced with permission from Elsevier. Box-and-whisker plots show median, IQR and range.

LETTER

https://doi.org/10.1038/s41586-018-0235-7

Altered exocrine function can drive adipose wasting in early pancreatic cancer

Laura V. Danai[1,13], Ana Babic[2,13], Michael H. Rosenthal[2], Emily A. Dennstedt[1], Alexander Muir[1], Evan C. Lien[1], Jared R. Mayers[1], Karen Tai[1], Allison N. Lau[1], Paul Jones-Sali[1], Carla M. Prado[3], Gloria M. Petersen[4], Naoki Takahashi[4], Motokazu Sugimoto[4], Jen Jen Yeh[5], Nicole Lopez[6], Nabeel Bardeesy[7], Carlos Fernandez-del Castillo[7], Andrew S. Liss[7], Albert C. Koong[8,9], Justin Bui[9,10], Chen Yuan[2], Marisa W. Welch[2], Lauren K. Brais[2], Matthew H. Kulke[2,11], Courtney Dennis[12], Clary B. Clish[12], Brian M. Wolpin[2]* & Matthew G. Vander Heiden[1,2,12]*

Malignancy is accompanied by changes in the metabolism of both cells and the organism[1,2]. Pancreatic ductal adenocarcinoma (PDAC) is associated with wasting of peripheral tissues, a metabolic syndrome that lowers quality of life and has been proposed to decrease survival of patients with cancer[3,4]. Tissue wasting is a multifactorial disease and targeting specific circulating factors to reverse this syndrome has been mostly ineffective in the clinic[5,6]. Here we show that loss of both adipose and muscle tissue occurs early in the development of pancreatic cancer. Using mouse models of PDAC, we show that tumour growth in the pancreas but not in other sites leads to adipose tissue wasting, suggesting that tumour growth within the pancreatic environment contributes to this wasting phenotype. We find that decreased exocrine pancreatic function is a driver of adipose tissue loss and that replacement of pancreatic enzymes attenuates PDAC-associated wasting of peripheral tissues. Paradoxically, reversal of adipose tissue loss impairs survival in mice with PDAC. When analysing patients with PDAC, we find that depletion of adipose and skeletal muscle tissues at the time of diagnosis is common, but is not associated with worse survival. Taken together, these results provide an explanation for wasting of adipose tissue in early PDAC and suggest that early loss of peripheral tissue associated with pancreatic cancer may not impair survival.

Models of autochthonous mice with PDAC recapitulate many features of human disease, including cachexia[7,8]. We confirmed the occurrence of severe adipose tissue and skeletal muscle wasting in advanced disease using two autochthonous mouse models with Kras(G12D) activation and loss of Trp53 function—either via *Trp53* deletion ($KP^{-/-}C$) or mutant $Trp53^{R172H}$ expression (KPC) (Fig. 1a–f). Because protein breakdown in tissues occurs early in PDAC[9], we assessed the kinetics of tissue wasting in the $KP^{-/-}C$ model and found a decrease in the mass of the adipose tissue by six weeks of age, whereas the weight of the pancreatic tissue was unchanged and plasma branched-chain amino acids (BCAAs), a measurement of peripheral tissue wasting[9], were modestly increased (Fig. 1g–i and Extended Data Fig. 1a). Adipose tissue wasting was greater than skeletal muscle wasting throughout disease progression (Fig. 1g, h), consistent with adipose tissue depletion occurring before skeletal muscle loss[10,11].

Histological analysis of gastrocnemius skeletal muscle in 6-week-old $KP^{-/-}C$ mice (early $KP^{-/-}C$) suggests muscle atrophy as characterized by a decrease in the cross-sectional area of myofibres (Extended Data Fig. 1b, c). Using micro-computed tomography imaging of the lower leg muscles (gastrocnemius skeletal muscle and soleus), we found a decreased muscle volume in early $KP^{-/-}C$ mice (Extended Data Fig. 1d, e). We also observed decreased muscle mass and elevated expression of genes involved in autophagy and ubiquitin–proteosome degradation, suggesting activation of muscle breakdown (Extended Data Fig. 1f, g). Histological analysis of adipose tissue also revealed smaller adipocytes and enhanced lipolysis in early $KP^{-/-}C$ mice (Extended Data Fig. 1h–l), suggestive of a reduction in triglyceride synthesis or triglyceride storage. These results are consistent with enhanced lipolysis[12,13] and muscle proteolysis[14] in advanced cancer; however, in $KP^{-/-}C$ mice these changes occur by six weeks of age, suggesting that peripheral tissue loss is initiated early in PDAC.

Although the $KP^{-/-}C$ model has predictable disease progression[9], there is widespread pancreatic transformation and cancer progresses in young mice such that tissue depletion could involve a failure to gain mass. To determine whether peripheral tissue wasting occurs in other PDAC models with a longer time until disease onset, we assessed adipose and skeletal muscle mass in mice in which oncogenic $Kras^{G12D}$ alone is activated in the pancreas (KC). KC mice develop premalignant lesions that progress to invasive tumours with advanced age[8]. In 15-week-old KC mice, an age preceding invasive cancer, we also observed decreased adipose and skeletal muscle mass (Fig. 1j–l and Extended Data Fig. 1m). These data indicate that peripheral tissue wasting in PDAC is initiated early and before frank cancer onset.

To test whether a PDAC-derived systemic factor mediates early tissue wasting in PDAC, we implanted PDAC cells isolated from autochthonous tumours either subcutaneously or orthotopically into the pancreas of syngeneic mice (Fig. 2a). When implanted at either location, solitary tumours form with similar histological features (Fig. 2b). Although mice were injected with the same number of cells, only the mice bearing orthotopic pancreatic tumours developed marked adipose wasting (Fig. 2c, d). Subcutaneous tumours are deficient in stromal components and may display reduced desmoplasia, although similar trichrome staining was observed for subcutaneous and orthotopic tumours, suggesting similar fibrosis (Fig. 2b). We isolated and immortalized pancreatic stellate cells (PSCs) from syngeneic mice[15] and injected PDAC cells alone or together with PSCs to form subcutaneous or pancreatic orthotopic tumours and tested whether this major component of the stroma contributes to peripheral tissue loss. Although we found that addition of PSCs did not alter peripheral tissue wasting, the location of the tumour still contributed to tissue wasting (Fig. 2e, f). We next analysed secreted factors that have been reported to be increased in late-stage PDAC and/or known to induce cachexia in other models[4,16] and found no differences between control and early $KP^{-/-}C$ mice (Extended Data Fig. 2). These results indicate that tumour growth in the pancreas, rather than

[1]Koch Institute for Integrative Cancer Research and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Dana-Farber Cancer Institute, Boston, MA, USA. [3]Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Alberta, Canada. [4]Mayo Clinic, Rochester, MN, USA. [5]Department of Surgery, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [6]University of California San Diego School of Medicine, La Jolla, CA, USA. [7]Massachusetts General Hospital Cancer Center, Harvard Medical School, Boston, MA, USA. [8]MD Anderson, Department of Radiation Oncology, Houston, TX, USA. [9]Stanford Cancer Institute, Stanford, CA, USA. [10]David Geffen School of Medicine at University of California, Los Angeles, CA, USA. [11]Section of Hematology/Oncology, Boston University and Boston Medical Center, Boston, MA, USA. [12]Broad Institute of MIT and Harvard University, Cambridge, MA, USA. [13]These authors contributed equally: Laura V. Danai, Ana Babic. *e-mail: bwolpin@partners.org; mvh@mit.edu

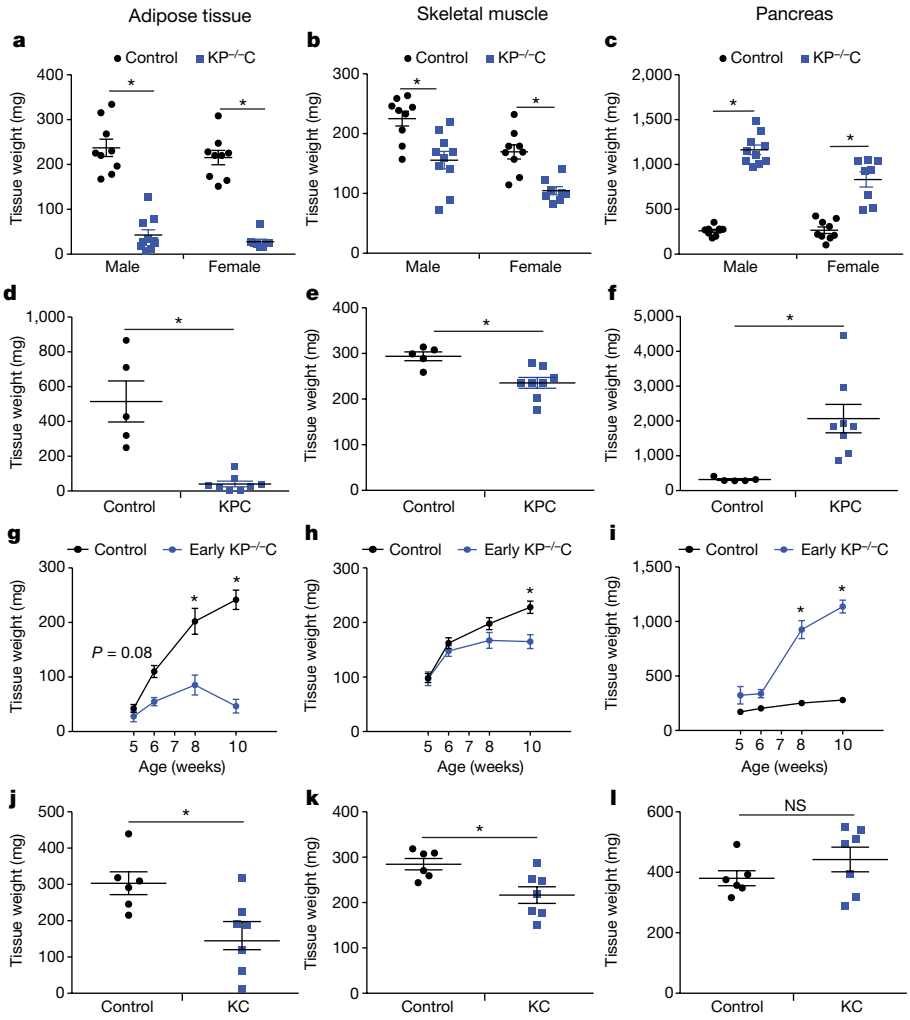


Fig. 1 | Early PDAC is associated with peripheral tissue wasting. a–c, Tissue weights of end-stage control and $KP^{-/-}C$ mice. $n = 9$ control and 10 $KP^{-/-}C$ male mice; $n = 9$ control and 8 $KP^{-/-}C$ female mice. **a**, Epididymal adipose tissue (males) and inguinal subcutaneous adipose tissue weight (females). $^*P < 1.25 \times 10^{-7}$ (for both comparisons). **b**, Gastrocnemius skeletal muscle weight. $^*P = 0.002$ (males), $^*P = 0.0003$ (females). **c**, Pancreas weight. $^*P < 1.16 \times 10^{-5}$ (for both comparisons). **d–f**, Tissue weights of end-stage male control and KPC mice. $n = 5$ control and 8 KPC mice. **d**, Epididymal adipose tissue weight. $^*P = 0.0003$. **e**, Gastrocnemius skeletal muscle weight. $^*P = 0.006$. **f**, Pancreas weight. $^*P = 0.007$. **g–i**, Peripheral tissue analysis during disease progression in control and $KP^{-/-}C$ male mice. Control and $KP^{-/-}C$ male mice of the following ages were used: 5 weeks ($n = 7$ and 5, respectively), 6 weeks ($n = 8$ and 7, respectively), 8 weeks ($n = 8$ and 10, respectively); 10 weeks ($n = 10$ and 9, respectively). **g**, Epididymal adipose tissue mass. Two-way ANOVA; 6 weeks, $^*P = 0.08$; 8 weeks, $^*P < 0.0001$; 10 weeks, $^*P < 0.0001$. **h**, Gastrocnemius skeletal muscle weight. Two-way ANOVA; 10 weeks, $^*P = 0.0007$. **i**, Pancreas weight. Two-way ANOVA; 6 weeks, $^*P = 0.07$; 8 weeks, $^*P < 0.0001$; 10 weeks, $^*P < 0.0001$. **j–l**, Tissue weights of 15-week-old male control and KC mice. $n = 6$ per group. **j**, Epididymal adipose tissue weight. $^*P = 0.01$. **k**, Gastrocnemius skeletal muscle weight. $^*P = 0.01$. **l**, Pancreas weight. NS, not significant; $P = 0.23$. Unless otherwise indicated, statistical analysis was performed using unpaired two-sided *t*-tests, data are mean ± s.e.m. and *n* represents the number of mice that were analysed.

a known systemic circulating PDAC-derived factor, promotes early wasting of the adipose tissue.

To investigate how tumour growth in the pancreas promotes adipose tissue wasting, we assessed systemic O_2 consumption, CO_2 production and calculated the respiratory exchange ratio. Both control and early $KP^{-/-}C$ mice displayed similar respiratory exchange ratios (Fig. 2g), arguing against a shift in whole-body fuel source utilization. However, O_2 consumption and CO_2 production were lower in early $KP^{-/-}C$ mice (Fig. 2h, i), suggesting decreased nutrient oxidation. Because food intake was similar between groups (Fig. 2j), early $KP^{-/-}C$ mice may metabolize less food and altered pancreatic function could explain these findings as well as the loss of adipose tissue.

To assess endocrine pancreatic function, we measured plasma glucagon and insulin. Whereas we found no significant differences in glucagon levels (Fig. 3a), early $KP^{-/-}C$ mice had lower levels of insulin in the fed state (Fig. 3b). To test for an insulin secretion defect, we measured plasma insulin after a bolus glucose injection and found no significant differences (Fig. 3c). These results indicate that the function of the endocrine pancreas was not altered in early PDAC; rather, impaired dietary absorption or starch breakdown may have caused the lower fed insulin levels. Consistent with this idea, we observed lower fed blood glucose levels in early $KP^{-/-}C$ mice (Fig. 3d). Furthermore, reduced insulin levels may promote increased lipolysis and adipose tissue loss.

To assess exocrine pancreatic function, we measured faecal lipid content and found higher faecal lipids in early $KP^{-/-}C$ mice (Fig. 3e). We also found decreased faecal protease activity and increased faecal protein content in these mice (Fig. 3f, g). Orthotopic implantation of PDAC cells into the pancreas also led to increased faecal protein content during disease progression (Fig. 3h). Decreased exocrine function of the pancreas could explain why early $KP^{-/-}C$ mice had decreased levels of O_2 consumption and CO_2 production despite normal food intake, leading to a starvation-like response with mobilization of energy stores from peripheral tissues. Indeed, mice bearing subcutaneous PDAC tumours that were fed a calorically restricted diet showed increased adipose tissue loss compared to skeletal muscle loss and reduced tumour size (Extended Data Fig. 3a–d).

To test whether decreased exocrine pancreatic function contributes to tissue wasting in PDAC, we supplemented a diet with pancreatic enzymes (Fig. 3i). Providing pancreatic enzymes attenuated adipose wasting in mice with PDAC (Fig. 3j). Furthermore, whereas mice with early PDAC displayed decreased fed glucose levels (Fig. 3d), mice with PDAC fed a diet supplemented with pancreatic enzymes displayed similar glucose levels to control littermates (Extended Data Fig. 3e). To control for potential food intake differences associated with adding pancreatic enzymes to a diet, we pair-fed mice to assure similar food consumption, and again observed attenuated adipose tissue loss in PDAC when providing pancreatic enzymes (Extended Data Fig. 3f–i). These results confirm that decreased pancreatic exocrine function mediates adipose tissue loss and contributes to peripheral tissue wasting in mice with early PDAC.

Cachexia has been proposed to worsen patient survival in various cancers including PDAC[3,4]. To determine whether adipose tissue wasting limits survival in PDAC, we assessed whether supplementation with pancreatic enzymes improved disease outcome. Despite reduced adipose tissue wasting, supplementation with pancreatic enzymes decreased survival of mice with PDAC (Fig. 3k), suggesting that peripheral tissue wasting may not always limit survival.

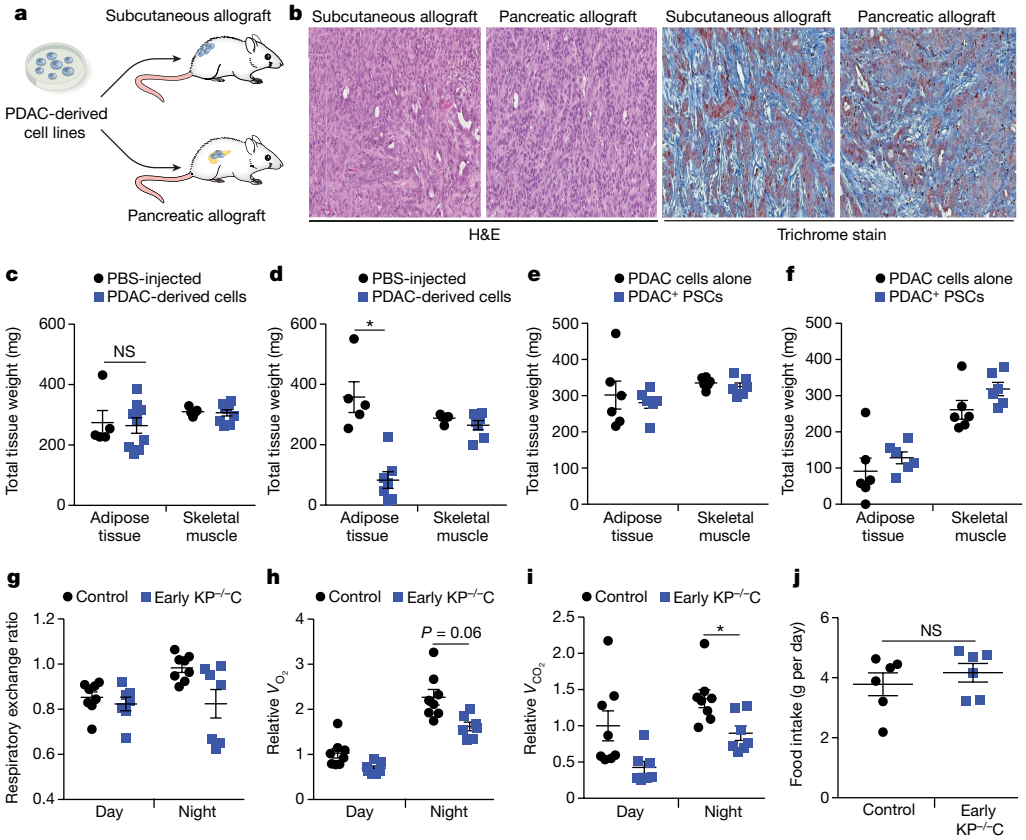


Fig. 2 | Pancreatic tumour growth promotes adipose tissue wasting. **a**, Schematic of experimental design. **b**, Representative haematoxylin and eosin (H&E) and Masson's trichrome histology of subcutaneous and orthotopic pancreatic tumours. $n = 4$. **c**, **d**, Epididymal adipose tissue and gastrocnemius skeletal muscle mass in mice injected with saline (PBS) or PDAC cells. **c**, Mice bearing subcutaneous allografts. $n = 5$ PBS-injected mice and 9 mice injected with PDAC cells. $P = 0.83$. **d**, Mice bearing pancreatic orthotopic allografts. $n = 5$ PBS-injected mice and 7 mice injected with PDAC cells. $P < 0.0001$. **e**, **f**, Epididymal adipose tissue and gastrocnemius skeletal muscle mass in mice injected with PDAC cells or PDAC cells together with PSCs. **e**, Mice bearing subcutaneous allograft tumours. $n = 6$ per group. $P = 0.63$. **f**, Mice bearing pancreatic allograft tumours. $n = 6$ per group. **g**–**j**, Metabolic cage measurements in male control (n = 8) and early $KP^{-/-}C$ ($n = 7$) mice. **g**. Respiratory exchange ratio. **h**, Relative volumetric oxygen consumption (V_{O_2}) calculated from area under the curve (AUC) measurements during the day and night. $P = 0.06$ (day) and $P = 0.129$ (night). **i**, Relative volumetric carbon dioxide release (V_{CO_2}) calculated from AUC measurements during the day and night. $P = 0.102$ (day) and $^*P = 0.01$ (night). **j**, Food intake (g consumed per day. $n = 6$ per group. $P = 0.45$. Unless otherwise indicated, statistical analysis was performed using unpaired two-sided *t*-tests, data are mean ± s.e.m. and *n* represents the number of mice that were analysed.

To investigate the association between peripheral tissue wasting and patient survival, we identified 782 patients at five US cancer centres with previously untreated PDAC, available clinical and outcome data, and banked blood samples (Extended Data Table 1). We quantified lumbar visceral and subcutaneous adipose tissue areas using pretreatment computed tomography scans[17,18] (Extended Data Fig. 4a). Although adipose tissue area was associated with multiple clinical factors (Extended Data Table 2), no association was found between adipose tissue wasting and patient survival in the full population (Table 1) or by disease stage (Extended Data Table 3a).

Studies of muscle wasting and survival of patients with PDAC have led to conflicting results[17,19–22], potentially because of differences in study design. We next used computed tomography imaging to measure the lumbar skeletal muscle index, a marker of muscle mass[17,19,20,22]. Using previously established cut points for sarcopenia[14], we found that 65% of patients displayed sarcopenia at diagnosis, and these patients did not have worse survival (Table 1). Notably, the prevalence of sarcopenia was not different between disease stages (localized, 64%; locally advanced, 70%; metastatic disease, 63%; $P = 0.40$). Because an interaction between sarcopenia and body mass index has been proposed[20,22], we evaluated whether sarcopenic obesity was associated with worse patient survival and observed no differences in survival between different patient groups (Table 1). Because optimal sarcopenia cut points are not well-defined[14], we investigated an agnostic skeletal muscle index classification with gender-specific quintiles and found no association with patient survival (Extended Data Table 3b). We also examined skeletal muscle area and attenuation and did not identify an association of these markers with patient survival (Extended Data Table 3b). Finally, because plasma BCAAs elevations reflect tissue wasting in early PDAC[9], we evaluated whether plasma BCAAs at diagnosis were associated with reduced survival and did not identify worse outcomes with elevated BCAAs (Extended Data Tables 4–6 and Extended Data Fig. 4b). Thus, in this multi-institutional patient population with newly diagnosed, previously untreated PDAC, we found no evidence that early skeletal muscle or adipose tissue wasting was associated with worse survival.

Independent of effects on survival, assessing peripheral tissue loss before overt disease onset may help to identify PDAC at earlier stages. In mouse models, we found that decreased exocrine function contributes to adipose tissue wasting. Many patients with PDAC experience loss of exocrine pancreatic function[23]; however, whether decreased exocrine function contributes to wasting in patients requires further study. Furthermore, correction of pancreatic exocrine function in mouse models reduces adipose tissue wasting but does not significantly affect muscle loss, suggesting that supplementation with pancreatic enzymes is either insufficient to fully restore the nutritional state or that additional factors contribute to muscle wasting. These other factors may include stroma-derived inflammatory signals that were not tested in this study. Nevertheless, the findings that a starvation-like state in mouse models of PDAC contributes to adipose tissue loss and increases

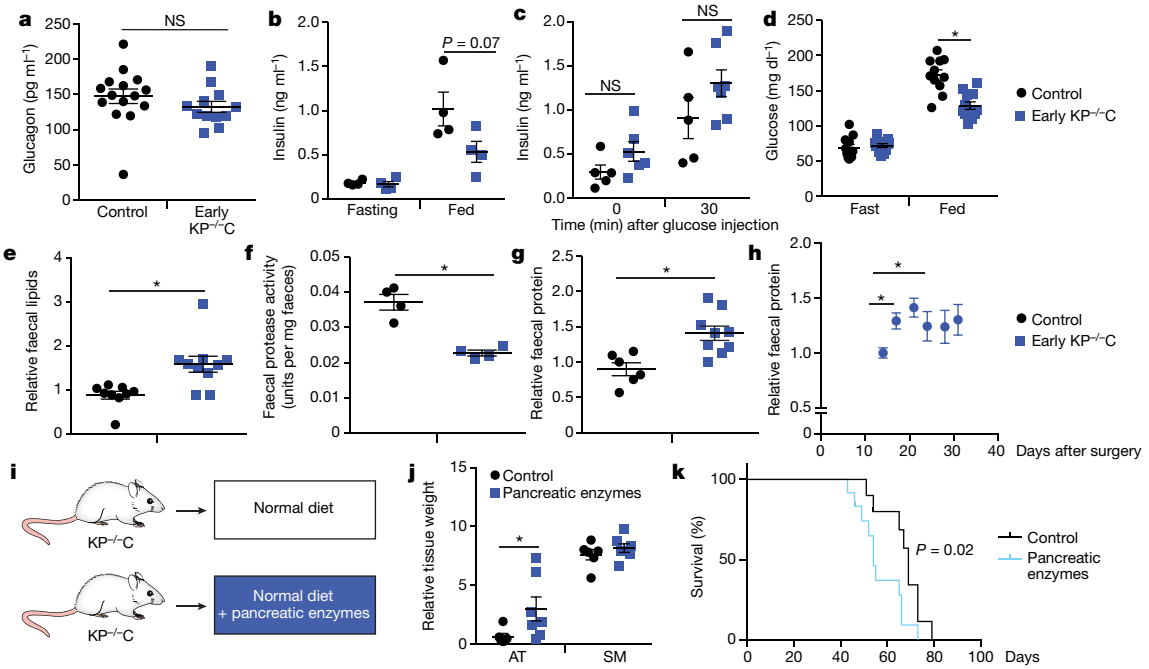


Fig. 3 | Decreased exocrine pancreatic function in early PDAC disease promotes tissue wasting. a–d, Endocrine function measurements in male control and early KP$^{-/-}$C mice. **a**, Fasted circulating glucagon levels. $n = 15$ control and 12 early KP$^{-/-}$C mice. $P = 0.28$. **b**, Insulin levels in overnight fasted and fed mice. $n = 4$ per group. **c**, Insulin levels in mice before and after a glucose injection. $n = 5$ control and 6 early KP$^{-/-}$C mice. $P = 0.13$ (time = 0) and $P = 0.16$ (time = 30). **d**, Glucose levels in overnight fasted and fed mice. $n = 11$ control and 12 early KP$^{-/-}$C mice. $^*P = 0.00008$. **e–g**, Faecal analysis of male control and early KP$^{-/-}$C mice. **e**, Total faecal lipid. $n = 9$ control and 10 early KP$^{-/-}$C mice. $^*P = 0.004$. **f**, Total faecal protease activity. $n = 4$ per group. $^*P = 0.0009$. **g**, Total faecal protein level. $n = 6$ control and 9 early KP$^{-/-}$C mice. $^*P = 0.004$. **h**, Total faecal protein levels over time in male C57BL/6J mice following orthotopic implantation of PDAC cells into the pancreas. $n = 5$. $P = 0.008$ (day 14 versus day 17) and $^*P = 0.005$ (day 14 versus day 21). **i**, Schematic of experimental design. **j**, Relative weights of KP$^{-/-}$C male mice fed a control diet or a diet supplemented with pancreatic enzymes. $n = 6$ control diet and 7 supplemented diet. $^*P = 0.033$. AT, adipose tissue; SM, skeletal muscle. **k**, Survival of male KP$^{-/-}$C mice fed the indicated diet. $n = 12$ per group. Mantel–Cox test, $P = 0.02$. Unless otherwise indicated, statistical analysis was performed using unpaired two-sided *t*-tests, data are mean ± s.e.m. and *n* represents the number of mice that were analysed.

Table 1 | Hazard ratios for death among cases with pancreatic cancer based on body composition

	Quintiles of body composition areas					
	1	2	3	4	5	***P* trend**[a]
Visceral adipose tissue area						
Number of cases	136	138	138	138	137	
Median (cm^2)	34.7	108.2	171.0	228.9	312.0	
Median overall survival (months)	12.3	10.2	11.5	11.3	11.1	
Hazard ratio (95% confidence interval)[b]	1.0	1.25 (0.96–1.62)	1.06 (0.82–1.38)	1.08 (0.83–1.41)	1.04 (0.79–1.36)	0.73
Hazard ratio (95% confidence interval)[c]	1.0	1.33 (1.01–1.74)	1.10 (0.82–1.47)	1.14 (0.84–1.56)	1.16 (0.82–1.63)	0.31
Subcutaneous adipose tissue area						
Number of cases	136	138	137	138	137	
Median (cm^2)	82.7	136.9	177.5	234.7	351.0	
Median overall survival (months)	11.9	10.6	12.0	11.6	10.2	
Hazard ratio (95% confidence interval)[b]	1.0	1.18 (0.91–1.53)	0.97 (0.75–1.26)	0.92 (0.71–1.19)	1.06 (0.81–1.38)	0.80
Hazard ratio (95% confidence interval)[c]	1.0	1.19 (0.91–1.57)	0.99 (0.74–1.32)	1.04 (0.77–1.41)	1.25 (0.86–1.82)	0.44
Sarcopenia[d]	**No**	**Yes**	***P* value**			
Number of cases[e]	248	462				
Median overall survival (months)	11.6	11.3				
Hazard ratio (95% confidence interval)[b]	1.0	1.03 (0.86–1.24)	0.74			
Hazard ratio (95% confidence interval)[c]	1.0	1.04 (0.85–1.27)	0.72			
Sarcopenia and obesity[d]	**Neither**	**Obese**	**Sarcopenic**	**Both**		
Number of cases[e]	54	186	191	262		
Median overall survival (months)	11.1	12.0	10.4	11.9		
Hazard ratio (95% confidence interval)[b]	1.0	0.98 (0.69–1.38)	1.11 (0.79–1.57)	0.95 (0.67–1.33)		
Hazard ratio (95% confidence interval)[c]	1.0	0.90 (0.60–1.35)	1.12 (0.80–1.58)	0.91 (0.63–1.32)		

[a]The two-sided *P* trend is calculated by entering the quintile-specific median value for adipose tissue area as a continuous variable in the Cox proportional hazards model.
[b]Cox proportional hazards model adjusted for age at diagnosis (continuous), gender (male or female), race (white, non-white or unknown), year of diagnosis (2000–2005, 2006–2010 or 2011–2015), institution (Dana-Farber/Brigham and Women's Cancer Center, Massachusetts General Hospital, Mayo Clinic, Stanford University or University of North Carolina) and cancer stage (local, locally advanced, metastatic or unknown).
[c]Cox proportional hazards model additionally adjusted for body mass index (continuous), diabetes history (none, ≤4 years, >4 years or unknown) and smoking status (never, past, current or unknown).
[d]Sarcopenia is defined as a skeletal muscle index (the ratio of the skeletal muscle area (in cm^2) to the height squared (in m^2)) of less than 55.4 cm^2 per m^2 for men and less than 38.9 cm^2 per m^2 for women. This was measured at baseline computed tomography imaging. Obesity is defined as a body mass index of more than 25 kg per m^2.
[e]For sarcopenia, 8 patients were excluded owing to missing information on height. For sarcopenia and obesity, a further 17 patients were excluded owing to missing information on weight.

survival are consistent with data that suggest that caloric restriction improves survival of mice with PDAC[24], perhaps via similar mechanisms. Although not examined here, changes in insulin levels may also contribute to survival differences upon reversal of a starvation-like state. The finding that pancreatic enzyme supplementation led to worse survival in mice also suggests that peripheral tissue wasting in early PDAC may be distinct from cachexia associated with late-stage disease. Nutritional intervention and pancreatic enzyme replacement are sometimes used in patients with PDAC[25] and a better understanding of the mechanisms that cause tissue wasting across cancer types and stages of disease is needed to design interventions that reduce functional disability and improve survival of patients with cancer.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0235-7.

Received: 28 February 2017; Accepted: 21 May 2018;
Published online: 20 June 2018

1. Koppenol, W. H., Bounds, P. L. & Dang, C. V. Otto Warburg's contributions to current concepts of cancer metabolism. *Nat. Rev. Cancer* **11**, 325–337 (2011).
2. Petruzzelli, M. & Wagner, E. F. Mechanisms of metabolic dysfunction in cancer-associated cachexia. *Genes Dev.* **30**, 489–501 (2016).
3. Dewys, W. D. et al. Prognostic effect of weight loss prior to chemotherapy in cancer patients. *Am. J. Med.* **69**, 491–497 (1980).
4. Mueller, T. C., Bachmann, J., Prokopchuk, O., Friess, H. & Martignoni, M. E. Molecular pathways leading to loss of skeletal muscle mass in cancer cachexia—can findings from animal models be translated to humans? *BMC Cancer* **16**, 75 (2016).
5. Fearon, K., Arends, J. & Baracos, V. Understanding the mechanisms and treatment options in cancer cachexia. *Nat. Rev. Clin. Oncol.* **10**, 90–99 (2013).
6. Penna, F. et al. Anti-cytokine strategies for the treatment of cancer-related anorexia and cachexia. *Expert Opin. Biol. Ther.* **10**, 1241–1250 (2010).
7. Flint, T. R. et al. Tumor-induced IL-6 reprograms host metabolism to suppress anti-tumor immunity. *Cell Metab.* **24**, 672–684 (2016).
8. Hingorani, S. R. et al. Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell* **4**, 437–450 (2003).
9. Mayers, J. R. et al. Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nat. Med.* **20**, 1193–1198 (2014).
10. Agustsson, T. et al. Mechanism of increased lipolysis in cancer cachexia. *Cancer Res.* **67**, 5531–5537 (2007).
11. Michaelis, K. A. et al. Establishment and characterization of a novel murine model of pancreatic cancer cachexia. *J. Cachexia Sarcopenia Muscle* **8**, 824–838 (2017).
12. Rydén, M. et al. Lipolysis—not inflammation, cell death, or lipogenesis—is involved in adipose tissue loss in cancer cachexia. *Cancer* **113**, 1695–1704 (2008).
13. Shaw, J. H. & Wolfe, R. R. Fatty acid and glycerol kinetics in septic patients and in patients with gastrointestinal cancer. The response to glucose infusion and parenteral feeding. *Ann. Surg.* **205**, 368–376 (1987).
14. Fearon, K. et al. Definition and classification of cancer cachexia: an international consensus. *Lancet Oncol.* **12**, 489–495 (2011).
15. Hwang, R. F. et al. Cancer-associated stromal fibroblasts promote pancreatic tumor progression. *Cancer Res.* **68**, 918–926 (2008).
16. Herrington, M. K., Arnelo, U. & Permert, J. On the role of islet amyloid polypeptide in glucose intolerance and anorexia of pancreatic cancer. *Pancreatology* **1**, 267–274 (2001).
17. Martin, L. et al. Cancer cachexia in the age of obesity: skeletal muscle depletion is a powerful prognostic factor, independent of body mass index. *J. Clin. Oncol.* **31**, 1539–1547 (2013).
18. Mourtzakis, M. et al. A practical and precise approach to quantification of body composition in cancer patients using computed tomography images acquired during routine care. *Appl. Physiol. Nutr. Metab.* **33**, 997–1006 (2008).
19. Choi, Y. et al. Skeletal muscle depletion predicts the prognosis of patients with advanced pancreatic cancer undergoing palliative chemotherapy, independent of body mass index. *PLoS ONE* **10**, e0139749 (2015).
20. Prado, C. M. et al. Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. *Lancet Oncol.* **9**, 629–635 (2008).
21. Rollins, K. E. et al. The impact of sarcopenia and myosteatosis on outcomes of unresectable pancreatic cancer or distal cholangiocarcinoma. *Clin. Nutr.* **35**, 1103–1109 (2016).
22. Tan, B. H., Birdsell, L. A., Martin, L., Baracos, V. E. & Fearon, K. C. H. Sarcopenia in an overweight or obese patient is an adverse prognostic factor in pancreatic cancer. *Clin. Cancer Res.* **15**, 6973–6979 (2009).
23. Vujasinovic, M., Valente, R., Del Chiaro, M., Permert, J. & Löhr, J. M. Pancreatic exocrine insufficiency in pancreatic cancer. *Nutrients* **9**, 183 (2017).
24. Lv, M., Zhu, X., Wang, H., Wang, F. & Guan, W. Roles of caloric restriction, ketogenic diet and intermittent fasting during initiation, progression and metastasis of cancer in animal models: a systematic review and meta-analysis. *PLoS ONE* **9**, e115147 (2014).
25. Laquente, B. et al. Supportive care in pancreatic ductal adenocarcinoma. *Clin. Transl. Oncol.* **19**, 1293–1302 (2017).

Acknowledgements We thank members of the Vander Heiden and Wolpin laboratories for discussions and the Koch Institute Swanson Biotechnology Center, particularly the Animal Imaging and Preclinical Testing Facility, for technical assistance. Major funding for this work was provided by the Lustgarten Foundation to B.M.W. and M.G.V.H. L.V.D. was supported by NIH Ruth Kirschstein Fellowship (F32CA210421). A.B. was supported by P50CA127003 and the Robert T. and Judith B. Hale Fund for Pancreatic Cancer Research. A.M. was supported by F32CA213810. E.C.L. was supported by the Damon Runyon Cancer Research Foundation (DRG-2299-17). A.N.L. is a Robert Black Fellow of the Damon Runyon Cancer Research Foundation (DRG-2241-15). B.M.W was supported by Robert T. and Judith B. Hale Fund for Pancreatic Cancer Research, NIH/NCI (U01CA210171), Department of Defense (CA130288), Pancreatic Cancer Action Network, Stand Up To Cancer, Noble Effort Fund, Peter R. Leavitt Family Fund, Wexler Family Fund, and Promises for Purple. M.G.V.H. was supported in part by a Faculty Scholar grant from the Howard Hughes Medical Institute, and acknowledges additional funding from Stand Up To Cancer, The Ludwig Center at MIT, the Koch Institute Frontier Awards, the MIT Center for Precision Cancer Medicine, and the NIH (R01CA168653, P30CA14051).

Reviewer information *Nature* thanks M. Löhr and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions L.V.D. designed, performed and analysed the animal experiments with input from M.G.V.H.; L.V.D., A.B., B.M.W. and M.G.V.H. wrote the manuscript with assistance from all other authors. E.A.D. and P.J.-S. assisted with animal experimentation. A.M. immortalized and A.N.L. isolated PSCs. E.C.L. performed caloric restriction experiments. J.R.M. performed muscle volume measurements and blood measurements. K.T. performed non-esterified fatty acid and glycerol assays. A.B., M.H.R., C.B.C. and B.M.W. designed the human study. C.M.P., G.M.P., N.T., M.S., J.J.Y., N.L., N.B., C.F.-d.C., A.S.L., A.C.K., J.B., C.Y., M.W.W., L.K.B., M.H.K. and B.M.W. were involved in patient recruitment and patient data collection. C.D. and C.B.C. were involved in metabolite measurements in patients. A.B. and M.H.R. analysed human data. B.M.W. supervised the human study.

Competing interests The authors declare no competing financial interests; however, M.G.V.H. discloses serving on the S.A.B. of Agios Pharmaceuticals and Aeglea Biotherapeutics.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0235-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0235-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to B.M.W. or M.G.V.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Cell culture. All cells were cultured in DMEM supplemented with 10% FBS (Corning) and 1% penicillin and streptomycin (Corning).

Western blot analysis. Antibodies recognizing pHSL (ser563) (4139) and total HSL (4107) were purchased from Cell Signaling Technologies.

Animal studies. All experiments performed in this study were approved by the MIT Committee on Animal Care (IACUC). Furthermore, for subcutaneous tumour growth, a maximum tumour burden of 2 cm^3 was permitted by our IACUC protocol and these limits were not exceeded. All mice in this study were fully backcrossed to the C57BL/6J background, and housed under a 12-h light and 12-h dark cycle, and cohoused with littermates with ad libitum access to water and food, unless otherwise stated. Furthermore, all experimental groups were age-matched and assigned based on genotype (or treatment). All animals were numbered and experiments were conducted in a blinded manner. After data collection, genotypes were revealed and animals assigned to groups for analysis. All measurements were collected from distinct animals, *n* represents biologically independent samples, and mice were analysed as tumours developed. No statistical methods were performed to predetermine sample size.

For studies using $Kras^{G12D}Trp53^{fl/fl}Pdx1^{cre}$ ($KP^{-/-}C$) mice, controls included littermates that lacked the Cre-recombinase allele, LSL-$Kras^{G12D}$ allele or both. For studies using $Kras^{G12D}Trp53^{R172H/+}Pdx1^{cre}$ (KPC) mice, control mice included littermates that lacked the LSL-$Kras^{G12D}$ allele or LSL-$p53^{172H}$ allele. For studies using $Kras^{G12D}Pdx1^{cre}$ (KC) mice, control mice included littermates expressing the Cre allele.

For subcutaneous or pancreatic orthotopic tumours, C57BL/6J mice (000664) were injected with 10^5 mouse PDAC cells isolated from C57BL/6J $KP^{-/-}C$ mice as previously described[9]. Phosphate-buffered saline (PBS) or PDAC cells were injected into the right flank (in 100 μl) or the pancreas (50 μl) of 8-week-old C57BL/6J mice. For protein faecal collection studies in mice bearing orthotopic pancreatic tumours, mice were injected with 10^5 mouse PDAC cells (50 μl) and allowed to recover from surgery for two weeks. Faecal matter was collected after this two-week recovery time every 2–3 days for approximately 2–3 weeks and placed in −20 °C until the day of analysis. For studies using PSCs, we isolated PSCs as previously described[26]. After a few passages, PSCs were immortalized and injected at a 1:1 ratio of PSCs to cancer cells as previously described[15].

For experiments using metabolic cages, 5–6-week-old mice were placed in metabolic cages and food intake, respiratory exchange ratio (RER), volumetric rate of O_2 consumption and volumetric of CO_2 production were measured over a three-day period (TSE Systems). RER was calculated using the following formula: $RER = V_{CO_2}/V_{O_2}$

For pairfeeding experiments, the amount of diet supplemented with pancreatic enzymes that was consumed per mouse per day was calculated for an average of five days. Animals with early PDAC were weighed, individually housed and randomly given a preweighed quantity of food that was either control diet or diet supplemented with pancreatic enzymes. Mice were weighed before the start of the food-pairing experiment to ensure both groups of mice had similar starting body weights.

For caloric-restriction studies, mice were injected subcutaneously (into both flanks) with PDAC-derived cells as described above. Animals were randomly placed on AIN-93 G (TD.94045) control diet or on the same diet at 40% restriction after tumours were palpable. Mice were individually housed and fed daily either 3.2 g per day (control mice) or 1.9 g per day (calorie-restricted mice) for a total of three weeks.

Animal diets. For diets supplemented with pancreatic enzymes, AIN-93G powdered diet (TD.94045) was purchased from Envigo and mixed with a commercial preparation of pancreatic enzymes as previously described[27].

Assessment of glucose metabolism. Glucose levels were measured using a Breeze-2 glucose meter (Bayer). For insulin measurements, we used an Ultrasensitive Mouse Insulin ELISA (#90080) following the manufacturer's protocol. For in vivo insulin secretion, mice were fasted for 16 h and intraperitoneally injected with glucose (1 g kg^{-1}), blood samples were collected and analysed at the indicated time points.

Ex vivo lipolysis. Ex vivo lipolysis assays of adipose explants were performed as previously described[28]. In brief, epididymal adipose tissues were collected and incubated at 37 °C, non-esterified fatty acids were measured using Wako diagnostics and glycerol release was measured using the Free Glycerol Determination kit (FG0100; Sigma-Aldrich) according to the manufacturer's instructions.

Hormone and metabolite measurements. Blood samples were collected in EDTA-containing tubes and centrifuged 3,000 r.p.m. for 15 min (4 °C). Circulating levels of IL-6 (M6000B, R&D Systems), corticosterone (80556, Crystal Chem), amylin (EIAM-AMY-1, RayBiotech) were measured according to the manufacturer's instructions. TNF, IFNγ, IL-10, IL-1β, IL-17 and IL-4 were measured using a Discovery Assay (Eve Technologies). Circulating BCAAs levels were measured as previously described[9].

Micro-computed tomography imaging in mice. All micro-computed tomography (CT) measurements in mice were performed using GE eXplore CT120. The scans were conducted at 70 kVp, 50 mA and 32 ms. There were 720 views, 0.5 degrees apart over a full 360-degree rotation. To assess muscle volume, a 3D Gaussian filter was used in an area extending from the right ankle to the proximal end of the fibula. This filtered dataset was used to determine areas corresponding to leg muscle while excluding other tissues. This was done in MATLAB using the connected components (bwconncomp) function. Having created the mask (segmented the leg muscle from other tissues), a histogram of the muscle region was calculated. Voxels falling within the density range of 160–200 Hounsfield units (HU) were considered muscle. This was done to correct for any overlap of the muscle mask with adjacent bone or adipose (which have higher and lower HU values, respectively).

Faecal assays. To assess total faecal protein, 10 mg of faeces was resuspended in lysis buffer (2% SDS, 150 mM NaCl, 0.5 M EDTA), sonicated and the protein concentration was assessed using a BCA assay according to the manufacturer's instructions. Total faecal protease activity was measured as previously described[29]. In brief, 10–30 mg of faecal matter was resuspended in 1 ml of buffer A (0.1%Triton X-100, 0.5 M NaCl, 100 mM $CaCl_2$), sonicated and centrifuged. The supernatant was then incubated with 3% Azo-Casein (Sigma-Aldrich, A2765) at 37 °C for 60 min. The reaction was stopped using 8% trichloroacetic acid and centrifuged. The absorbance of the supernatant (measured at 366 nm) was measured using a spectrophotometer. Total faecal lipids were measured as previously described[30]. In brief, 1,000 mg of faeces was collected and lipids were extracted using a 2:1 chloroform:methanol solution. The lipid fraction was dried using a stream of gaseous nitrogen and the vials were weighed.

Histological analysis. Tissues were fixed overnight with neutral-buffered 10% formalin, paraffin-embedded, sectioned and stained with haematoxylin and eosin or stained with Masson's trichrome using standard protocols.

Statistics for animal data. All graphs were generated using Prism (GraphPad) software and data are mean ± s.e.m. Unless otherwise indicated, *P* values were determined using unpaired two-sided *t*-tests. Statistical outliers were measured using a Grubb's outlier test (Prism) and excluded from the final analysis.

Human population study. Our study population included patients with pancreatic cancer from five US cancer centres: Dana-Farber/Brigham and Women's Cancer Center (DF/BWCC), Massachusetts General Hospital (MGH), Mayo Clinic, Stanford University and University of North Carolina-Chapel Hill (UNC). We included 782 patients with pancreatic adenocarcinoma who were diagnosed between 2000 and 2015, and had a stored plasma sample collected before receiving any treatment for their malignancy, including surgery, radiation or chemotherapy. A total of 778 patients at the five institutions met these criteria and had a plasma sample collected within 30 days before their pathological diagnosis and 60 days after this diagnosis. Of these patients, 687 had a CT scan performed during this time period for the analysis of body composition. The overall study was approved by the Dana-Farber/Harvard Cancer Center IRB, and data abstraction and blood sample collection was approved by each individual institutional IRB. All participants provided informed consent.

Human plasma samples and metabolite profiling. Blood was collected in sterile EDTA tubes, and processed within 3 h (Dana-Farber Cancer Institute, Mayo Clinic, Stanford University and UNC) or 24 h (MGH) for separation into plasma and other components. Plasma was aliquoted into cryovials and stored at −80 °C. Plasma samples were thawed once on wet ice to aliquot into smaller sample volumes for analysis, de-identified and sent for analysis in a single shipment.

Liquid chromatography tandem mass spectrometry (LC–MS) analyses were conducted using a Shimadzu Nexera X2 U-HPLC (Shimadzu) coupled to a Q Exactive hybrid quadrupole orbitrap mass spectrometer (Thermo Fisher Scientific). Polar metabolites were extracted from plasma (10 μl) using 90 μl of 74.9:24.9:0.2 v/v/v acetonitrile:methanol:formic acid containing stable isotope-labelled internal standards (valine-d_8 (Sigma-Aldrich) and phenylalanine-d_8 (Cambridge Isotope Laboratories)). The samples were centrifuged (10 min, 9,000*g*, 4 °C) and the supernatants were injected directly onto a 150 × 2-mm^2, 3-μm Atlantis HILIC column (Waters). The column was eluted isocratically at a flow rate of 250 μl min^{-1} with 5% mobile phase A (10 mM ammonium formate and 0.1% formic acid in water) for 0.5 min followed by a linear gradient to 40% mobile phase B (acetonitrile with 0.1% formic acid) over 10 min. MS analyses were carried out using electrospray ionization in the positive ion mode using full scan analysis over 70–800 *m*/*z* at 70,000 resolution and 3-Hz data acquisition rate. Other MS settings were: sheath gas 40, sweep gas 2, spray voltage 3.5 kV, capillary temperature 350 °C, S-lens RF 40, heater temperature 300 °C, microscans 1, automatic gain control target 1×10^6, and maximum ion time 250 ms. Raw data were processed using TraceFinder software (Thermo Fisher Scientific). Metabolite identities were confirmed using authentic reference standards.

To evaluate reproducibility of BCAA measurements, we included 77 blinded quality control plasma samples within the larger sample set, from eight plasma

quality control pools. The mean coefficients of variance were 7.6% for isoleucine, 8.0% for leucine and 7.3% for valine. Because blood samples from MGH were processed after >3 h, we evaluated the reproducibility of BCAA measurements between samples processed at different time points after collection. For 10 patients, we processed blood immediately and at 24 h and measured plasma BCAAs. Spearman correlation coefficients for the BCAA levels at the two time points were 0.95 for isoleucine, 0.95 for leucine and 0.92 for valine. All blood samples were collected into EDTA plasma tubes, except for 28 blood samples collected as serum from MGH patients. To evaluate BCAAs in plasma versus serum tubes, we measured BCAA levels for 10 patients who had simultaneous collection of plasma and serum. Spearman correlation coefficients for plasma BCCAs by plasma versus serum samples were 0.84 for isoleucine, 0.76 for leucine and 0.94 for valine. Given the high correlation coefficients for plasma BCAAs for both time of blood processing and plasma versus serum collection, we included all MGH patients in our study population.

Computed tomography imaging quantification of muscle and adipose tissue in patients. Muscle, visceral adipose tissue and subcutaneous adipose tissue areas were measured through the third lumbar vertebra landmark on CT imaging as previously published[17,18]. Scans from the DF/BWCC, MGH, Stanford University and UNC study sites ($n = 363$) were manually segmented using Slice-O-Matic software (v.4.3; Tomovision) by trained image analysts with final verification from a board-certified radiologist. All paraspinal and abdominal wall muscles were included in the muscle area. The Hounsfield CT attenuation scale was used to constrain adipose tissue and muscle selection. The Hounsfield scale is a linear transformation of the linear attenuation coefficient, with fixed points calibrated for air at −1,000 HU and water at 0 HU. All medical CT scanners are routinely calibrated to this scale. Pixel attenuation constraints from −29 to 150 HU were used for muscle and from −180 to −30 HU for adipose tissue as previously published[17,18]. The visceral adipose tissue compartment was defined by the peritoneum; all extraperitoneal adipose tissue was included in the subcutaneous compartment. Pixel dimensions were extracted from scanner parameters embedded within the scan data; the total area was measured as the product of segmented pixel count and pixel area. We calculated skeletal muscle index (SMI) as the ratio of skeletal muscle area (cm^2) to height squared (m^2). Muscle attenuation was defined as the average Hounsfield attenuation of all pixels in the muscle area. Analyst performance was tested, and a test–retest coefficient of variation <1.1% was observed for all analysts and parameters.

Scans from the Mayo Clinic site ($n = 324$) were analysed using in-house developed software written in MATLAB (MATLAB 2015b, MathWorks) with manual correction by a trained image analyst. The software automatically fits three concentric closed contours at the air–skin boundary, the subcutaneous adipose–muscle boundary and the abdominal wall–peritoneal adipose boundary using hierarchical morphological classification constrained by a prior probability shape model. These boundaries defined the same zones as were described for the Slice-O-Matic method. The software also automatically created masks for bone and colonic content and these masks were used to exclude bone and colonic content from being included as muscle or adipose tissue. The final boundaries were verified by a board-certified radiologist. Adipose and muscle areas were then calculated for each compartment using the same attenuation constraints as described for the Slice-O-Matic method.

To ensure consistency across sites, we analysed 20 cases using both methods and found a coefficient of variation of 2.4%, 1.7% and 3.8% between the methods for the skeletal muscle, subcutaneous adipose tissue and visceral adipose tissue areas, respectively. The maximum observed differences were 4.1%, 3.6%, and 15.2% for these compartments.

Covariate data and statistical analysis of patient data. Using patient questionnaires and medical records, we extracted information on patient and clinical characteristics, including: age at diagnosis, gender, race/ethnicity, height, weight at blood collection, diabetes history (no diabetes, diabetes of duration ≤4 years, diabetes of duration >4 years), tobacco use (never, past, current), primary tumour location (head or uncinate, body, tail, other), cancer stage (local, locally advanced, metastatic), year of diagnosis and survival time.

Body composition measurements and plasma BCAAs by patient characteristics were compared using Wilcoxon rank-sum test or Kruskal–Wallis test. A Pearson correlation test was used to evaluate the linear relationship between body composition measurements, plasma BCAAs and other patient characteristics. To evaluate the associations between patient survival, body composition measurements, and plasma BCAAs, we used multivariable-adjusted Cox proportional hazards regression and calculated hazard ratios and 95% confidence intervals. Survival time was calculated as the time from diagnosis to death or date of last follow-up if the patient was still alive. In an initial multivariate model, we adjusted for age at diagnosis (continuous), gender (male or female), race (white, non-white or unknown), year of diagnosis (2000–2005, 2006–2010 or 2011–2015), institution (DF/BWCC, MGH, Mayo Clinic, Stanford University or UNC), and cancer stage (localized, locally advanced, metastatic or unknown). In a second multivariate model, we additionally adjusted for body mass index (BMI) (continuous), history of diabetes (none, ≤4 years, >4 years or unknown), and smoking status (never, past, current or unknown), which have previously been shown to associate with patient survival[31,32].

Because cut-offs to define clinically relevant body composition categories are not well-defined, we also divided patients into quintiles by adipose tissue and muscle tissue measurements. Because the distribution of body composition measurements differed by gender (Extended Data Table 2), we created gender-specific quintiles for all body composition measurements. *P* values for linear trends were calculated by entering the quantile-specific median value for body composition measurements in the Cox proportional hazards models.

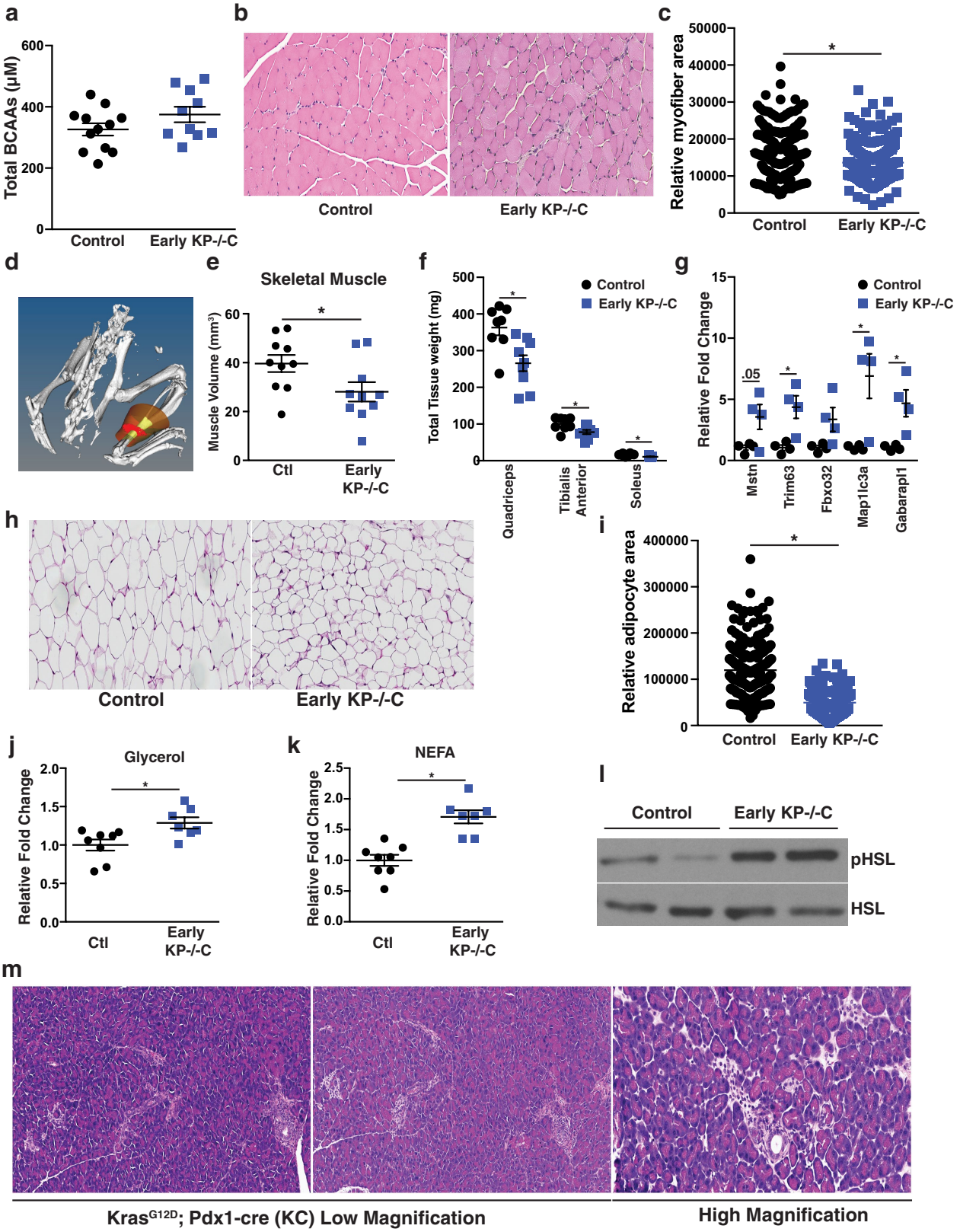
To evaluate sarcopenia, we used the SMI criteria established by a recent international consensus definition of cancer cachexia[14]. The SMI cut-offs were 55.4 cm^2 per m^2 for men and 38.9 cm^2 per m^2 for women. To evaluate sarcopenic obesity, we generated a combined variable, for which overweight or obesity status is defined as BMI ≥ 25 kg per m^2, similar to a previous study in patients with PDAC[22]. The resulting four categories were overweight/obese and sarcopenic, overweight/obese and non-sarcopenic, non-overweight/obese and sarcopenic, non-overweight/obese and non-sarcopenic. For plasma BCAAs, patients were divided into quartiles, with quartiles 2–4 compared to quartile 1 as the reference. *P* values for trends were calculated by entering quartile-specific median values for total plasma BCAAs as a continuous variable in the Cox proportional hazards model.

We performed stratified analyses by disease stage and statistical interactions were assessed using the Wald test of the cross-product term between stage (localized, locally advanced, metastatic or unknown) and tissue measurements (continuous). We also calculated the hazard ratio for each institution and measured heterogeneity across centres using the Cochran's *Q* statistic[33]. This statistic is a weighted sum of squared deviations of the estimate of an individual study from the overall estimate obtained by meta-analysis. All analyses were performed with SAS 9.2 statistical analysis software, and all *P* values are from two-sided tests.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

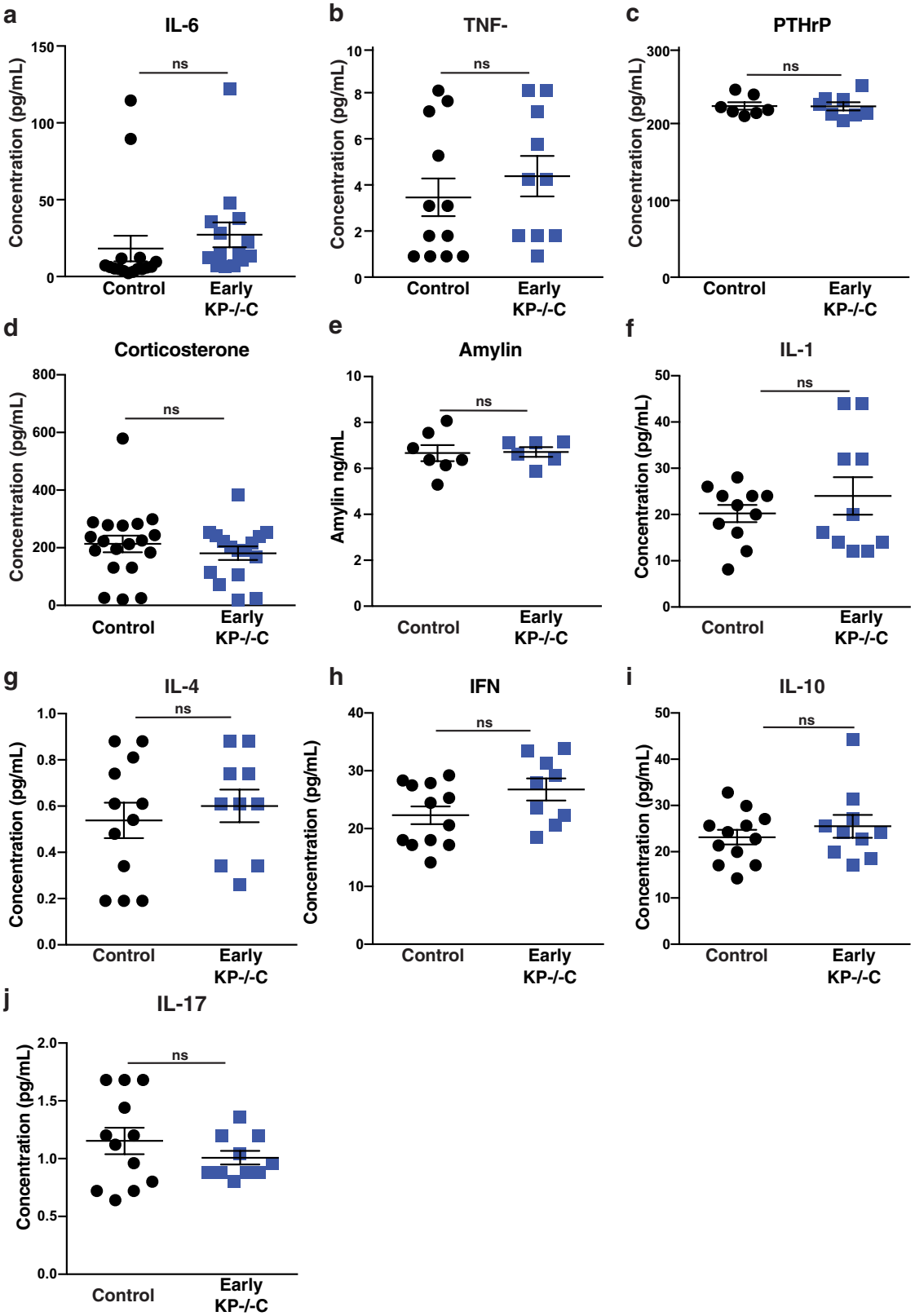
Data availability. Source Data for the graphical representations found in Figs. 1–3 and Extended Data Figs. 1–3 can be found in the online version of the paper. All other data that support the findings of this study are available upon request from the corresponding author.

26. Apte, M. Isolation of quiescent pancreatic stellate cells from rat and human pancreas. *Pancreapedia: Exocrine Pancreas Knowledge Base* https://doi.org/10.3998/panc.2011.10 (2011).
27. Loftus, S. K. et al. Acinar cell apoptosis in *Serpini2*-deficient mice models pancreatic insufficiency. *PLoS Genet.* **1**, e38 (2005).
28. DiStefano, M. T. et al. The lipid droplet protein hypoxia-inducible gene 2 promotes hepatic triglyceride deposition by inhibiting lipolysis. *J. Biol. Chem.* **290**, 15175–15184 (2015).
29. Gabbi, C. et al. Pancreatic exocrine insufficiency in LXRβ$^{-/-}$ mice is associated with a reduction in aquaporin-1 expression. *Proc. Natl Acad. Sci. USA* **105**, 15052–15057 (2008).
30. Kraus, D., Yang, Q. & Kahn, B. B. Lipid extraction from mouse feces. *Bio Protoc.* **5**, e1375 (2015).
31. Yuan, C. et al. Survival among patients with pancreatic cancer and long-standing or recent-onset diabetes mellitus. *J. Clin. Oncol.* **33**, 29–35 (2015).
32. Yuan, C. et al. Prediagnostic body mass index and pancreatic cancer survival. *J. Clin. Oncol.* **31**, 4229–4234 (2013).
33. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).

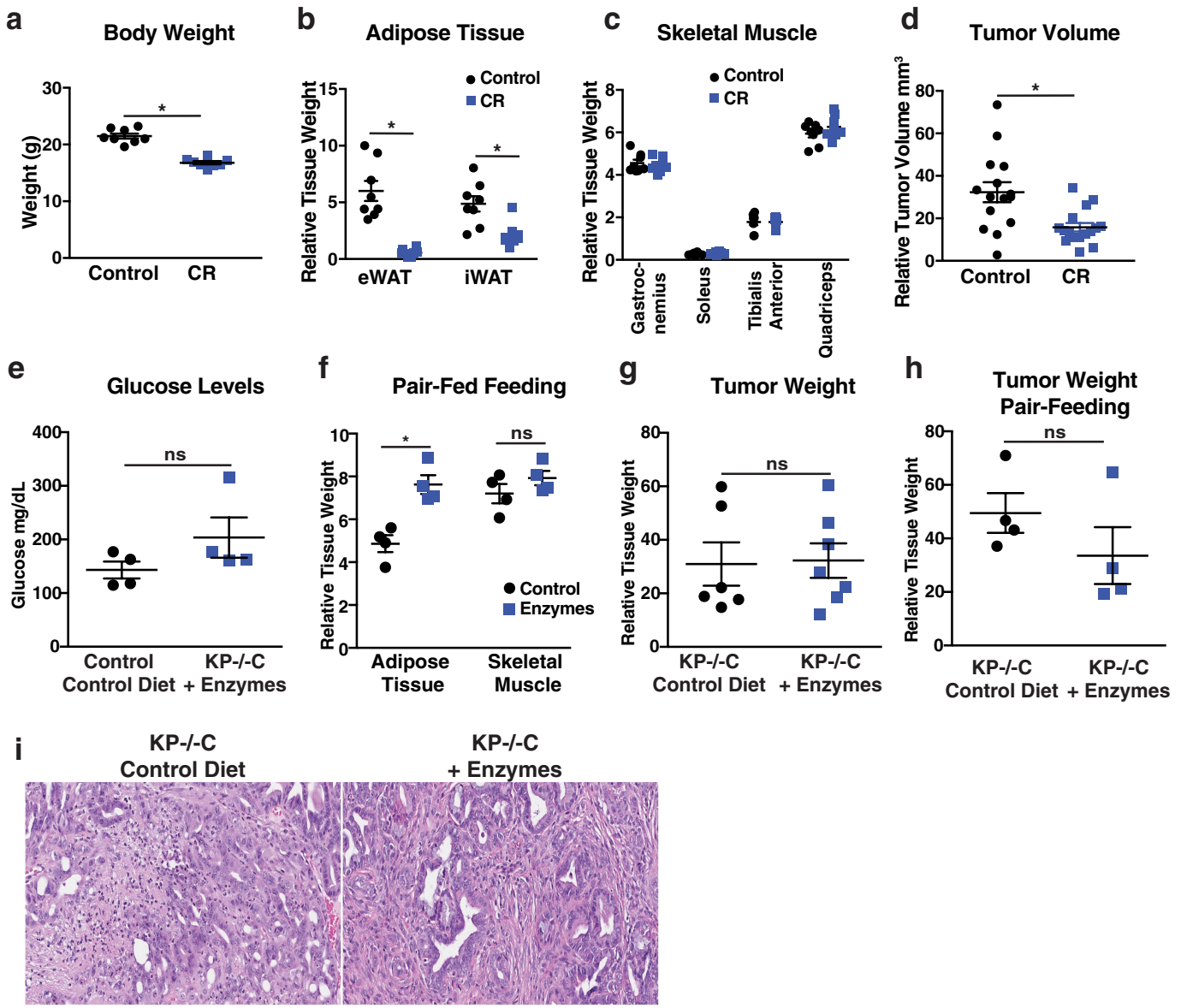


Extended Data Fig. 1 | See next page for caption.

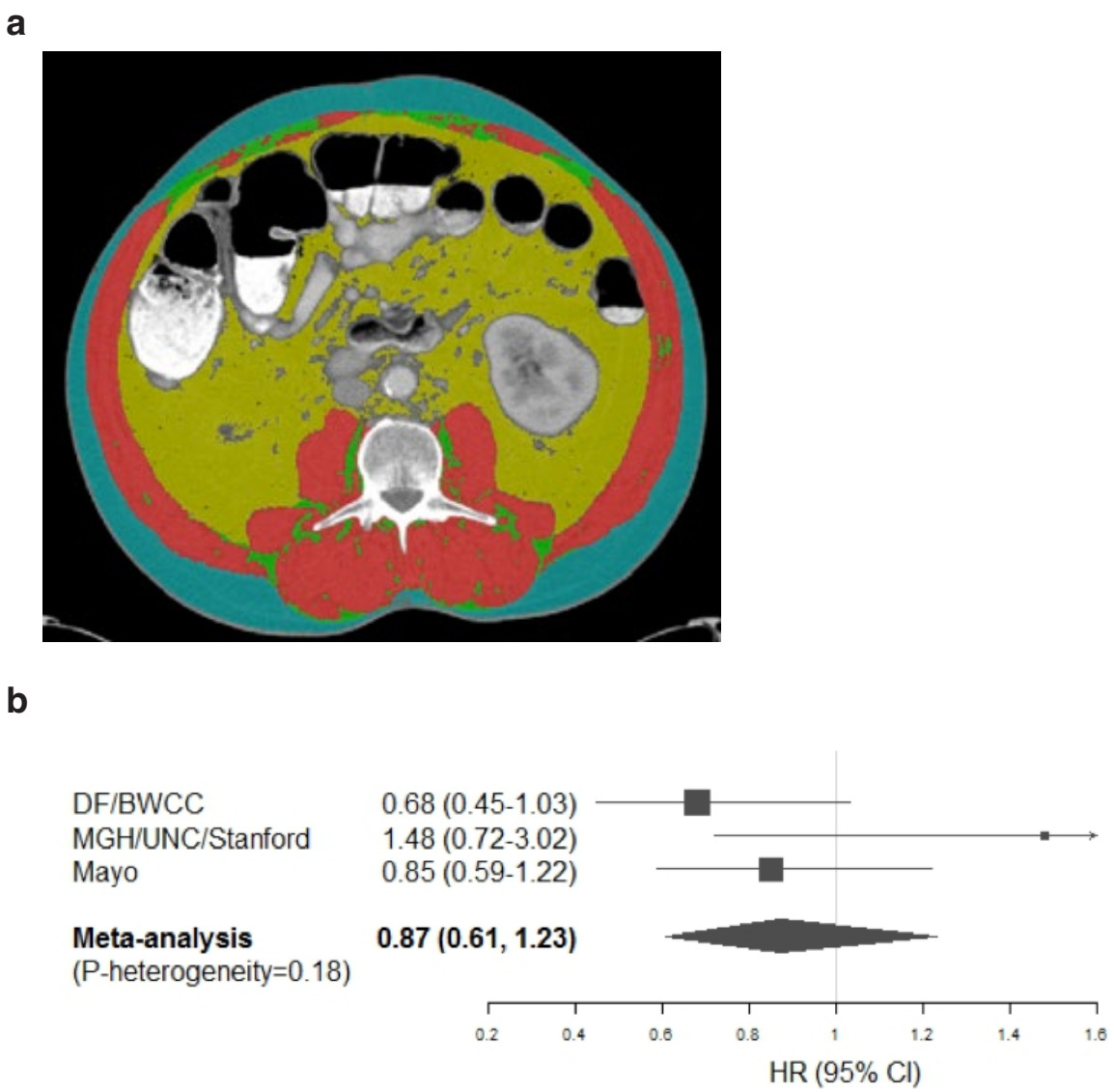
Extended Data Fig. 1 | PDAC is associated with adipose and skeletal muscle wasting. a, Circulating BCAAs (valine, leucine, and isoleucine) in male control ($n = 12$) and early $KP^{-/-}C$ ($n = 10$) mice. **b**, Representative histology of H&E-stained gastrocnemius skeletal muscle of control and early $KP^{-/-}C$ mice. $n = 4$ per group. **c**, Relative myofibre area in male control and early $KP^{-/-}C$ mice. $n = 3$ per group. $^*P < .0001$. **d**, Representative 3D μCT imaging reconstruction of soleus and gastrocnemius skeletal muscle (highlighted in red). **e**, Relative soleus and gastrocnemius skeletal muscle as assessed by micro-CT scan of control and early $KP^{-/-}C$ male mice. $n = 10$ per group. $P = 0.04$. **f**, Skeletal muscle tissue mass of the indicated muscle groups in male control ($n = 8$) and early $KP^{-/-}C$ mice ($n = 9$). $^*P = 0.006$ (quadriceps), $^*P = 0.02$ (tibialis anterior), $^*P = 0.004$ (soleus). **g**, Relative mRNA expression of the indicated genes assessed by RT–qPCR. $n = 4$ per group. $P = 0.05$ (*Mstn*), $^*P = 0.01$ (*Trim63*), $^*P = 0.07$ (*Fbxo32*), $^*P = 0.00004$ (*Map1lc3a*), $^*P = .006$ (*Gabarapl1*). **h**, Representative histology of H&E-stained epididymal adipose tissue from control and early $KP^{-/-}C$ male mice. $n = 4$ per group. **i**, Relative adipocyte area in male control and early $KP^{-/-}C$ mice. $n = 3$ per group. $P < 0.0001$. **j**, Glycerol release in ex vivo adipose tissue explants from control and early $KP^{-/-}C$ male mice. $n = 7$ per group. $P = 0.01$. **k**, Non-esterified fatty acid (NEFA) release in ex vivo adipose tissue explants from control ($n = 8$) and $KP^{-/-}C$ male ($n = 7$) mice. $P = 0.0002$. **l**, Representative western blot analysis of phosphorylated (p)HSL and HSL expression in adipose tissue of control and early $KP^{-/-}C$ male mice. $n = 3$ per group. **m**, Representative H&E histology images of the pancreas of 15-week-old KC male mice. $n = 5$ per group. Unless otherwise indicated, statistical analysis was performed using unpaired two-sided *t*-tests, data are mean ± s.e.m. and *n* represents the number of mice that were analysed.



Extended Data Fig. 2 | Systemic circulating factors are not altered in early PDAC. a–j, Circulating levels of the indicated factors in control and early $KP^{-/-}C$ mice. **a**, IL-6. $n = 16$ control and 14 early $KP^{-/-}C$ mice. ns, not significant; $P = 0.45$. **b**, TNF-α. $n = 12$ control and 10 early $KP^{-/-}C$ mice. $P = 0.45$. **c**, PTHrP. $n = 7$ control and 8 early $KP^{-/-}C$ mice. $P = 0.94$. **d**, Corticosterone. $n = 19$ control and 16 early $KP^{-/-}C$ mice. $P = 0.40$. **e**, Amylin. $n = 7$ control and 6 early $KP^{-/-}C$ mice. $P = 0.91$. **f**, IL-1β. $n = 11$ control and 10 early $KP^{-/-}C$ mice. $P = 0.39$. **g**, IL-4. $n = 12$ control and 10 early $KP^{-/-}C$ mice. $P = 0.56$. **h**, IFNγ. $n = 12$ control and 9 early $KP^{-/-}C$ mice. $P = 0.08$. **i**, IL-10. $n = 12$ control and 10 early $KP^{-/-}C$ mice. $P = 0.42$. **j**, IL-17. $n = 12$ and 10 early $KP^{-/-}C$ mice $P = 0.29$. Unless otherwise indicated, statistical analysis was performed using unpaired two-sided *t*-tests, data are mean ± s.e.m. and *n* represents the number of mice that were analysed.



Extended Data Fig. 3 | Decreased exocrine pancreatic function in early PDAC disease promotes adipose tissue loss. a–d, C57BL/6J mice bearing PDAC-derived subcutaneous tumours fed a control diet or the same diet at 40% caloric restriction (CR) for 3 weeks. $n = 8$ per group. **a**, Body weight. $^*P < 0.0001$. **b**, Epididymal adipose tissue (eWAT) and inguinal adipose tissue (iWAT) mass normalized to body weight. $^*P < 0.0001$ for epididymal adipose tissue and $^*P = 0.0034$ for inguinal adipose tissue. **c**, Skeletal muscle mass of the indicated muscle groups normalized to body weight. **d**, Tumour volume normalized to body weight. $^*P = 0.002$. $n = 15$ tumours from 8 control mice (2 tumours per mouse; for one mouse one of the tumours did not grow) and $n = 16$ tumours from 8 calorie-restricted mice. **e**, Fed plasma glucose levels of 7-week-old male control and $KP^{-/-}C$ mice that were fed the indicated diets. $n = 4$ per group. $P = 0.18$. **f**, Tissue weights normalized to body weight in $KP^{-/-}C$ mice pair-fed indicated diets. $n = 4$ per group. $^*P = 0.01$. ns, $P = 0.53$. **g**, **h**, Tumour weight normalized to body weight of $KP^{-/-}C$ mice that were fed the indicated diets for 1 week. **g**, $KP^{-/-}C$ mice were fed the indicated diets. $n = 6$ control and 7 enzyme-supplemented. $P = 0.9$. **h**, $KP^{-/-}C$ mice were pair-fed the indicated diets. $n = 4$ per group. $P = 0.26$. **i**, Representative histology of H&E-stained autochthonous pancreatic tumours of $KP^{-/-}C$ mice that were fed the indicated diets. $n = 4$ per group. Unless otherwise indicated, statistical analysis was performed using unpaired two-sided *t*-tests, data are mean ± s.e.m. and *n* represents the number of mice that were analysed.



Extended Data Fig. 4 | Use of CT imaging to assess patient body composition and relationship between plasma BCAA levels and patient survival by study site. a, Representative CT image used to analyse body composition. Skeletal muscle is shown in red, intramuscular adipose tissue is shown in green, visceral adipose tissue is shown in yellow and subcutaneous adipose tissue is shown in blue. **b**. Hazard ratios (HRs) and 95% confidence intervals (CI) for the association between plasma BCAAs and patient survival, comparing the top and bottom quartile, calculated using Cox proportional hazards model adjusted for age at diagnosis (continuous), gender (male or female), race (white, non-white or unknown), year of diagnosis (2000–2005, 2006–2010 or 2011–2015), cancer stage (local, locally advanced, metastatic or unknown), BMI (continuous), diabetes history (none, ≤4 years, >4 years or unknown) and smoking status (never, past, current or unknown). The pooled hazard ratios were calculated using the DerSimonian and Laird random-effects model. The solid squares and horizontal lines correspond to the study site-specific multivariate hazard ratio and 95% confidence interval, respectively. The area of the solid square reflects the study site-specific weight (inverse of the variance). The filled diamond represents the pooled hazard ratio and 95% confidence interval. The solid vertical line indicates a hazard ratio of 1.0. $n = 778$.

Extended Data Table 1 | Characteristics for patients with pancreatic cancer

a.

	Overall	DF/BWCC	MGH	Stanford	UNC	Mayo
No. of patients	782	281	58	36	57	350
Diagnosis period, years						
2000-2005	110 (14)	25 (9)	0 (0)	14 (39)	0 (0)	71 (20)
2006-2010	292 (37)	90 (32)	18 (31)	15 (42)	24 (42)	145 (41)
2011-2015	380 (49)	166 (59)	40 (69)	7 (19)	33 (58)	134 (38)
Age at diagnosis, years	66.0 (10.7)	64.3 (10.0)	69.4 (10.2)	67.9 (13.1)	68.2 (11.1)	66.2 (10.7)
Female sex	347 (44)	133 (48)	28 (48)	11 (32)	31 (54)	144 (41)
White race	698 (89)	252 (90)	39 (67)	24 (67)	39 (68)	344 (98)
Body mass index, kg/m^2	27.4 (5.2)	26.7 (5.2)	27.0 (5.7)	24.7 (4.7)	27.4 (4.7)	28.2 (5.0)
Diabetes history						
No diabetes	565 (72)	194 (69)	46 (79)	28 (78)	32 (56)	265 (76)
Diabetes ≤4 years	100 (13)	38 (14)	5 (9)	4 (11)	8 (14)	45 (13)
Diabetes >4 years	54 (7)	20 (7)	4 (7)	2 (6)	7 (12)	21 (6)
Unknown	63 (8)	29 (10)	3 (5)	2 (6)	10 (18)	19 (5)
Tobacco use						
Never	335 (43)	127 (45)	14 (24)	18 (50)	20 (35)	156 (45)
Past	330 (42)	121 (43)	36 (62)	13 (36)	25 (44)	135 (39)
Current	68 (9)	30 (11)	8 (14)	4 (11)	12 (21)	14 (4)
Missing	49 (6)	3 (1)	0 (0)	1 (3)	0 (0)	45 (13)
Primary tumor location						
Head / Uncinate	481 (62)	161 (57)	41 (71)	24 (67)	49 (86)	206 (59)
Body	142 (18)	64 (23)	6 (10)	8 (22)	3 (5)	61 (17)
Tail	97 (12)	47 (17)	11 (19)	3 (8)	5 (9)	31 (9)
Body and tail	31 (4)	5 (2)	0 (0)	1 (3)	0 (0)	25 (7)
Head and body	22 (3)	0 (0)	0 (0)	0 (0)	0 (0)	22 (6)
Other	9 (1)	4 (1)	0 (0)	0 (0)	0 (0)	5 (1)
Cancer stage						
Local	258 (33)	52 (19)	51 (88)	5 (14)	54 (95)	96 (27)
Locally advanced	157 (20)	47 (17)	0 (0)	21 (58)	1 (2)	88 (25)
Metastatic	308 (39)	175 (62)	4 (7)	9 (25)	2 (4)	118 (34)
Unknown	59 (8)	7 (2)	3 (5)	1 (3)	0 (0)	48 (14)
Median survival, months						
Local	20.6	20.5	25.8	6.3	17.4	22.4
Locally advanced	11.0	13.0	n/a	7.9	1.4	11.1
Metastatic	7.0	6.9	15.5	4.9	3.9	8.2

b.

	Full Study Population (Pt No.)	**Blood Samples** (Pt No.)	**CT Images** (Pt No.)
DF/BWCC	281	280	244
MGH	58	58	38
Stanford	36	34	33
UNC	57	56	48
Mayo Clinic	350	350	324
Total	**782**	**778**	**687**

a, Characteristics for each study site. **b**, Number of patients for which blood samples and CT images were available. Continuous variables are reported as mean ± s.d.; categorical variables are reported as number (percentage) at time of diagnosis unless otherwise noted.

Extended Data Table 2 | Body composition characteristics for patients with pancreatic cancer

	Pt No.	SMI (cm^2/m^2)	Muscle area (cm^2)	Muscle attenuation (HU)	Subcutaneous fat area (cm^2)	Visceral fat area (cm^2)
Age						
< 60 yrs	194	47.9 (9.9)	144.4 (37.4)	39.1 (9.4)	206.9 (116.6)	155.0 (107.9)
60 - <70 yrs	238	46.6 (11.1)	134.2 (34.9)	36.9 (9.6)	214.7 (115.1)	163.2 (103.5)
≥ 70 yrs	255	41.7 (7.7)	118.4 (28.9)	32.4 (9.4)	185.3 (94.2)	163.7 (111.0)
P-value[a]		$9.8x10^{-13}$	$3.3x10^{-14}$	$3.0x10^{-12}$	0.03	0.53
Gender						
Male	384	49.5 (8.8)	153.8 (27.6)	37.0 (9.4)	185.8 (102.2)	203.0 (111.9)
Female	303	39.5 (8.5)	102.6 (17.8)	34.5 (10.3)	221.5 (113.8)	107.8 (72.7)
P-value[a]		$3.7x10^{-51}$	$6.5x10^{-88}$	0.004	$2.3x10^{-6}$	$2.1x10^{-31}$
Race						
White	621	45.0 (10.0)	131.4 (34.4)	35.6 (9.8)	201.2 (105.5)	163.9 (105.5)
Non-White	42	46.8 (11.1)	129.7 (39.8)	37.6 (10.7)	216.4 (137.3)	117.6 (118.3)
P-value[a]		0.24	0.60	0.15	0.80	0.001
Diabetes						
No diabetes	498	44.4 (9.4)	129.4 (35.1)	37.0 (9.7)	191.4 (99.6)	148.8 (105.6)
Diabetes ≤4 yrs	92	45.8 (8.5)	133.5 (33.0)	33.3 (10.1)	224.4 (112.3)	189.2 (99.0)
Diabetes >4 yrs	48	48.0 (10.6)	144.1 (35.9)	35.1 (8.7)	216.5 (130.4)	207.2 (116.1)
P-value[a]		0.01	0.01	0.004	0.03	$3.2x10^{-6}$
BMI, kg/m^2						
<18.5	14	35.1 (4.6)	92.4 (15.0)	44.9 (12.2)	61.3 (51.9)	24.3 (20.7)
18.5-24.9	220	41.0 (8.0)	117.8 (31.2)	39.0 (9.2)	135.6 (64.8)	89.9 (67.2)
>24.9-29.9	260	45.6 (8.1)	134.7 (32.1)	35.4 (8.7)	193.6 (73.3)	176.0 (91.1)
>29.9-34.9	117	49.4 (10.2)	146.4 (36.3)	32.9 (9.7)	264.3 (83.0)	228.1 (100.2)
>35	54	50.7 (10.0)	147.8 (35.6)	29.2 (10.9)	398.4 (121.2)	269.1 (128.0)
P-value[a]		$3.6x10^{-22}$	$7.1x10^{-19}$	$2.3x10^{-12}$	$1.0x10^{-29}$	$1.1x10^{-29}$
Current smoking						
No	586	45.2 (10.1)	131.3 (35.1)	35.8 (9.8)	201.6 (105.6)	163.0 (109.4)
Yes	55	42.9 (8.6)	121.0 (29.9)	35.7 (11.1)	183.5 (112.6)	123.8 (88.4)
P-value[a]		0.11	0.05	0.81	0.07	0.02
Cancer stage						
Local	213	45.3 (10.6)	130.7 (32.9)	34.6 (9.5)	217.5 (113.3)	170.7 (110.9)
Locally advanced	147	45.1 (9.9)	132.6 (35.6)	37.8 (10.3)	186.5 (103.3)	152.5 (99.6)
Metastatic	279	45.3 (9.7)	131.6 (36.0)	35.9 (9.5)	190.5 (98.6)	155.6 (106.3)
Unknown	48	43.5 (9.3)	126.9 (34.5)	35.2 (11.3)	241.0 (142.0)	176.0 (120.3)
P-value[a]		1.00	0.87	0.01	0.02	0.28
Primary tumor location						
Head/Uncinate	423	44.6 (10.2)	129.0 (33.8)	35.5 (10.1)	205.0 (110.7)	159.4 (105.6)
Body	131	45.0 (9.0)	130.6 (32.5)	36.9 (9.2)	185.8 (94.4)	155.9 (104.3)
Tail	78	46.5 (9.8)	137.0 (38.0)	34.5 (9.9)	213.9 (112.2)	177.9 (122.7)
P-value[a]		0.22	0.26	0.10	0.17	0.51

Body composition measurements reported as mean ± s.d.
[a]Two-sided *P* values were calculated using Wilcoxon rank-sum tests.

Extended Data Table 3 | Hazard ratios (with 95% confidence intervals) for death among cases of pancreatic cancer by body composition measurements using computed tomography

a.

	Pt. No.	Extreme Quintiles[a,b]	Per S.D.[a]	P-interaction[c]
Visceral fat area (cm^2)				0.59
Localized	213	2.02 (0.97-4.21)	1.15 (0.88-1.50)	
Locally advanced	147	0.53 (0.22-1.32)	0.81 (0.64-1.04)	
Metastatic	279	0.98 (0.54-1.76)	0.97 (0.80-1.17)	
Subcutaneous fat area (cm^2)				0.72
Localized	213	1.29 (0.58-2.87)	1.44 (1.05-1.98)	
Locally advanced	146	0.64 (0.26-1.60)	0.83 (0.68-1.00)	
Metastatic	279	0.99 (0.54-1.80)	0.91 (0.73-1.12)	
SMI (cm^2/m^2)				0.90
Localized	210	0.66 (0.30-1.46)	0.84 (0.64-1.12)	
Locally advanced	145	0.59 (0.27-1.29)	0.87 (0.70-1.08)	
Metastatic	277	1.54 (0.85-2.79)	1.05 (0.84-1.30)	
Muscle area (cm^2)				0.47
Localized	213	0.82 (0.34-1.98)	0.86 (0.62-1.19)	
Locally advanced	147	0.40 (0.16-1.00)	0.81 (0.62-1.06)	
Metastatic	279	1.56 (0.77-3.18)	1.10 (0.82-1.45)	
Muscle attenuation (HU)				0.49
Localized	210	0.32 (0.16-0.67)	0.74 (0.60-0.90)	
Locally advanced	147	0.99 (0.50-1.96)	0.94 (0.75-1.19)	
Metastatic	275	1.01 (0.60-1.69)	0.96 (0.80-1.16)	

b.

	Quintiles of Body Composition Measurements					*P*-trend[f]
	1	2	3	4	5	
SMI (cm^2/m^2)						
Median OS (mo.)	11.1	11.0	10.8	11.6	11.7	
Hazard ratio (95% CI)[d]	1.0	0.98 (0.76-1.27)	1.04 (0.80-1.36)	0.92 (0.70-1.19)	0.91 (0.68-1.21)	0.35
Hazard ratio (95% CI)[e]	1.0	1.01 (0.78-1.31)	1.04 (0.79-1.37)	0.94 (0.71-1.24)	0.93 (0.68-1.27)	0.47
Muscle area (cm^2)						
Median OS (mo.)	10.2	12.1	10.8	11.6	11.3	
Hazard ratio (95% CI)[d]	1.0	0.75 (0.58-0.98)	0.90 (0.69-1.17)	0.81 (0.62-1.05)	0.84 (0.63-1.12)	0.50
Hazard ratio (95% CI)[e]	1.0	0.78 (0.60-1.02)	0.91 (0.69-1.20)	0.82 (0.62-1.10)	0.87 (0.64-1.20)	0.71
Muscle attenuation (HU)						
Median OS (mo.)	10.2	11.2	10.8	12.7	11.7	
Hazard ratio (95% CI)[d]	1.0	0.89 (0.68-1.15)	0.92 (0.70-1.20)	0.75 (0.57-0.98)	0.87 (0.65-1.15)	0.16
Hazard ratio (95% CI)[e]	1.0	0.86 (0.65-1.13)	0.89 (0.67-1.18)	0.71 (0.53-0.96)	0.80 (0.57-1.11)	0.10

a, Patients were stratified by cancer stage at diagnosis. **b**, All patients. $n = 687$ patients.
[a]Cox proportional hazards model adjusted for age at diagnosis (continuous), gender (male or female), race (white, non-white or unknown), year of diagnosis (2000–2005, 2006–2010 or 2011–2015), institution (DF/BWCC, MGH, Mayo Clinic, Stanford University or UNC), BMI (continuous), diabetes history (none, ≤4 years, >4 years or unknown) and smoking status (never, past, current or unknown).
[b]Comparing the highest and lowest quintile.
[c]Two-sided *P* interaction calculated by Wald test of the cross-product between body component measurements (continuous) and stage (local, locally advanced, metastatic, unknown).
[d]Cox proportional hazards model adjusted for age at diagnosis (continuous), gender (male or female), race (white, non-white or unknown), year of diagnosis (2000–2005, 2006–2010 or 2011–2015), institution (DF/BWCC, MGH, Mayo Clinic, Stanford University or UNC) and cancer stage (local, locally advanced, metastatic or unknown).
[e]Cox proportional hazards model adjusted for age at diagnosis (continuous), gender (male or female), race (white, non-white or unknown), year of diagnosis (2000–2005, 2006–2010 or 2011–2015), institution (DF/BWCC, MGH, Mayo Clinic, Stanford University or UNC), cancer stage (local, locally advanced, metastatic or unknown), BMI (continuous), diabetes history (none, ≤4 years, >4 years or unknown) and smoking status (never, past, current or unknown).
[f]Two-sided *P*-trend values calculated by entering the quintile-specific median value for body composition measurements as a continuous variable in a Cox proportional hazards model.

Extended Data Table 4 | Plasma BCAA levels and clinical characteristics of pancreatic cancer cases

	Pt No.	Total BCAAs (μM)
Age, years		
< 60	220	351.6 (174.9)
60 - <70	271	327.7 (104.3)
≥ 70	287	297.7 (84.8)
P-value[a]		$1.3x10^{-6}$
Gender		
Male	432	346.1 (101.2)
Female	346	295.0 (144.1)
P-value[a]		$3.4x10^{-18}$
Race		
White	696	322.2 (98.3)
Non-White	45	290.6 (122.2)
P-value[a]		0.10
Diabetes		
No diabetes	563	322.3 (129.0)
Diabetes ≤4 years	99	330.9 (113.8)
Diabetes >4 years	54	341.8 (122.7)
P-value[a]		0.24
Body-mass index, kg/m^2		
<18.5	14	286.3 (120.4)
18.5-24.9	245	316.7 (168.0)
>24.9-29.9	295	322.6 (91.5)
>29.9-34.9	131	337.0 (111.7)
>35	63	335.3 (99.0)
P-value[a]		0.01
Current smoking		
No	661	323.0 (101.2)
Yes	68	328.8 (291.0)
P-value[a]		0.04
Cancer stage		
Local	256	306.6 (89.0)
Locally advanced	155	321.1 (96.1)
Metastatic	308	332.9 (110.0)
Unknown	59	348.1 (283.8)
P-value[a]		0.05
Primary tumor location		
Head/Uncinate	479	311.7 (137.8)
Body	141	336.4 (107.1)
Tail	96	345.5 (70.1)
P-value[a]		$6.6x10^{-7}$

Plasma BCAA measurements reported as mean ± s.d.
[a]Two-sided *P* values calculated using Wilcoxon rank-sum test.

Extended Data Table 5 | Pearson correlation coefficients for BCAAs, body composition measurements and patient characteristics

Variable	BMI	Age at diagnosis	Valine	Leucine	Isoleucine	Total BCAA	SMI	Muscle area	Muscle attenuation	Subcutaneous fat area	Visceral fat area
BMI	1.00	-0.14[a]	0.10[a]	0.06	0.01	0.06	0.43[b]	0.38[b]	-0.30[b]	0.74[b]	0.58[b]
Age at diagnosis		1.00	-0.19[b]	-0.17[b]	-0.14[a]	-0.18[b]	-0.29[b]	-0.33[b]	-0.32[b]	-0.09[a]	0.03
Valine			1.00	0.92[b]	0.78[b]	0.95[b]	0.22[b]	0.29[b]	0.18[b]	0.03	0.18[b]
Leucine				1.00	0.91[b]	0.99[b]	0.17[b]	0.24[b]	0.17[b]	-0.01	0.12[a]
Isoleucine					1.00	0.93[b]	0.10[a]	0.14[a]	0.13[a]	-0.01	0.07
Total BCAA						1.00	0.17[b]	0.24[b]	0.17[b]	0.01	0.13[a]
SMI							1.00	0.84[b]	0.20[b]	0.15[a]	0.37[b]
Muscle area								1.00	0.23[b]	0.08[a]	0.44[b]
Muscle attenuation									1.00	-0.40[b]	-0.34[b]
Subcutaneous fat area										1.00	0.43[b]
Visceral fat area											1.00

$n = 687$ patients.
[a]Two-sided $P < 0.05$.
[b]Two-sided $P < 0.0001$.

Extended Data Table 6 | Hazard ratios (with 95% confidence intervals) for death among pancreatic cancer cases by plasma BCAA levels at diagnosis

	Quartiles of Plasma Branched Chain Amino Acids				*P*-trend[a]
	1	2	3	4	
Total BCAAs					
Median (μM)	217.7	284.3	343.3	425.0	
Range (μM)	72.9-255.7	255.8-312.8	312.9-374.3	374.5-2327.2	
Median OS (mo.)	10.6	10.8	12.4	12.1	
Hazard ratio (95% CI)[b]	1.00	0.95 (0.76-1.19)	0.79 (0.62-0.99)	0.81 (0.64-1.02)	0.04
Hazard ratio (95% CI)[c]	1.00	0.96 (0.76-1.20)	0.78 (0.62-0.99)	0.82 (0.64-1.04)	0.06
Isoleucine					
Median (μM)	43.0	54.9	66.1	86.0	
Range (μM)	12.7-49.8	49.9-60.4	60.5-73.6	73.6-824.7	
Median OS (mo.)	10.8	11.7	11.6	11.6	
Hazard ratio (95% CI)[b]	1.0	0.89 (0.71-1.12)	0.93 (0.74-1.16)	0.89 (0.71-1.12)	0.42
Hazard ratio (95% CI)[c]	1.0	0.89 (0.70-1.12)	0.90 (0.72-1.14)	0.90 (0.71-1.14)	0.49
Leucine					
Median (μM)	73.6	99.9	121.8	152.5	
Range (μM)	23.6-89.0	89.3-108.0	108.3-132.6	132.7-844.6	
Median OS (mo.)	10.6	11.1	12.1	11.9	
Hazard ratio (95% CI)[b]	1.0	0.91 (0.73-1.14)	0.81 (0.64-1.02)	0.82 (0.65-1.04)	0.08
Hazard ratio (95% CI)[c]	1.0	0.93 (0.74-1.16)	0.82 (0.65-1.03)	0.84 (0.66-1.07)	0.12
Valine					
Median (μM)	98.6	129.0	155.5	191.3	
Range (μM)	35.4-114.0	114.2-141.9	142.0-171.3	171.4-658.0	
Median OS (mo.)	9.9	11.5	12.4	12.1	
Hazard ratio (95% CI)[b]	1.0	0.78 (0.62-0.99)	0.71 (0.56-0.90)	0.71 (0.56-0.91)	0.01
Hazard ratio (95% CI)[c]	1.0	0.79 (0.62-1.00)	0.71 (0.56-0.90)	0.74 (0.57-0.94)	0.02

$n = 778$ patients.
[a]Two-sided *P*-trend calculated by entering the quartile-specific median value for individual plasma BCAAs as a continuous variable in a Cox proportional hazards model.
[b]Cox proportional hazards model adjusted for age at diagnosis (continuous), gender (male or female), race (white, non-white or unknown), year of diagnosis (2000–2005, 2006–2010 or 2011–2015), institution (DF/BWCC, MGH, Mayo, Stanford University or UNC) and cancer stage (local, locally advanced, metastatic or unknown).
[c]Cox proportional hazards model additionally adjusted for BMI (continuous), diabetes history (none, ≤4 years, >4 years or unknown) and smoking status (never, past, current or unknown).

LETTER

https://doi.org/10.1038/s41586-018-0238-4

A naturally occurring antiviral ribonucleotide encoded by the human genome

Anthony S. Gizzi[1,5], Tyler L. Grove[1,5]*, Jamie J. Arnold[2], Joyce Jose[2], Rohit K. Jangra[3], Scott J. Garforth[1], Quan Du[1], Sean M. Cahill[1], Natalya G. Dulyaninova[1], James D. Love[4], Kartik Chandran[3], Anne R. Bresnick[1], Craig E. Cameron[2] & Steven C. Almo[1,4]*

Viral infections continue to represent major challenges to public health, and an enhanced mechanistic understanding of the processes that contribute to viral life cycles is necessary for the development of new therapeutic strategies[1]. Viperin, a member of the radical *S*-adenosyl-L-methionine (SAM) superfamily of enzymes, is an interferon-inducible protein implicated in the inhibition of replication of a broad range of RNA and DNA viruses, including dengue virus, West Nile virus, hepatitis C virus, influenza A virus, rabies virus[2] and HIV[3,4]. Viperin has been suggested to elicit these broad antiviral activities through interactions with a large number of functionally unrelated host and viral proteins[3,4]. Here we demonstrate that viperin catalyses the conversion of cytidine triphosphate (CTP) to 3′-deoxy-3′,4′-didehydro-CTP (ddhCTP), a previously undescribed biologically relevant molecule, via a SAM-dependent radical mechanism. We show that mammalian cells expressing viperin and macrophages stimulated with IFNα produce substantial quantities of ddhCTP. We also establish that ddhCTP acts as a chain terminator for the RNA-dependent RNA polymerases from multiple members of the Flavivirus genus, and show that ddhCTP directly inhibits replication of Zika virus in vivo. These findings suggest a partially unifying mechanism for the broad antiviral effects of viperin that is based on the intrinsic enzymatic properties of the protein and involves the generation of a naturally occurring replication-chain terminator encoded by mammalian genomes.

Consideration of genome context shows that in vertebrates viperin is always immediately adjacent to a gene annotated as cytidylate monophosphate kinase 2 (*CMPK2*) and in some organisms, such as *Lacinutrix mariniflava*, these two genes are fused (Extended Data Fig. 1a). These observations suggested that viperin might modify a nucleotide. Evaluation of approximately 200 constructs derived from 8 mammalian species identified a *Rattus norvegicus* viperin (rVIP), including residues 51–361, which exhibited excellent properties in solution (Extended Data Fig. 1b, c). We screened rVIP against a diverse set of nucleotides and deoxynucleotides, looking for enhanced 5′-deoxyadenosine (5′-dA) formation as an indicator of substrate activation (See Supplementary Information for details). As with most other radical-SAM proteins, when provided with dithionite as an artificial electron donor rVIP performs reductive cleavage of SAM in the absence of substrates. As shown in Fig. 1a, CTP selectively activates 5′-dA production by approximately 130-fold relative to protein alone. Liquid chromatography shows that, in addition to 5′-dA (9.1 min), another product is present (5.2 min) (Fig. 1b), which exhibits a UV-visible spectrum similar to CTP ($\lambda_{max} = 271$ nm), indicating that the pyrimidine ring is not notably modified by the viperin-mediated reaction (Fig. 1c). Liquid chromatography coupled with mass spectrometry (LC–MS) showed the new compound exhibited a negative-ion mass to charge ratio ($-m/z$) of 464.1, which is 18 Da less than the $-m/z$ of 482.1 of CTP (Fig. 1d, e).

^{13}C–^{13}C correlation spectroscopy (COSY) NMR on uniformly labelled $^{13}C_9{}^{15}N_3$-viperin product, and ^{1}H–^{13}C 2D heteronuclear single quantum coherence (2D HSQC) chemical-shift analysis and ^{31}P NMR analysis on natural abundance viperin product were performed (see Supplementary Information for details, Extended Data Figs. 1d, 2; see Supplementary Table 1 for chemical-shift data). Taken together, all available NMR and mass spectrometry data are consistent with the rVIP-catalysed conversion of CTP to ddhCTP, in which dehydration—involving the loss of both the 4′ hydrogen and the 3′ hydroxyl group—occurs without rearrangement of the carbon skeleton. It is notable that MoaA, which catalyses the conversion of GTP to (8S)-3′,8-cyclo-7,8-dihydroguanosine 5′-triphosphate[5], is the enzyme with the highest sequence similarity to viperin for which an unambiguous functional annotation exists. Additionally, the recent report of the structure of viperin from *Mus musculus* proposed that the presumptive catalytic site shares high amino acid conservation with MoaA and may operate on a nucleotide-like substrate[6]. These observations are consistent with the above data.

rVIP has a K_m of 183 ± 28 μM for CTP and produces ddhCTP with a maximum velocity of 0.185 ± 0.007 min^{-1} (Fig. 1f). The intracellular concentration of CTP typically falls in the 1 mM range, which agrees well with the K_m obtained for rVIP for CTP[7], and the rate of ddhCTP formation is consistent with that of other radical-SAM enzymes with their native substrates[8]. 5′-dA and ddhCTP production is tightly coupled, with one molecule of 5′-dA generated for every ddhCTP (Extended Data Fig. 3a). rVIP also produces ddhCTP when the reaction is initiated by an enzymatic reducing system, indicating that dithionite does not direct an unanticipated side reaction between rVIP and CTP (Extended Data Fig. 3b).

A recent report described a radical-SAM enzyme from the thermophilic fungus *Thielavia terrestris* (58% sequence identity to human viperin) that was capable of catalysing the coupling of UDP–glucose and 5′-dA to generate an uncharacterized product with a *m/z of* 818.1[9]. In addition, a preliminary report suggested that viperin homologues from *Methanofollis liminatans* (archaea, 35% sequence identity to human viperin) and *Trichoderma virens* (fungi, 55% identity) catalyse the radical-based condensation of 5′-dA and isopentenyl pyrophosphate to yield adenylated isopentyl pyrophosphate[10]. Substrate activation and competition assays demonstrate that mammalian viperin does not catalyse these transformations (Fig. 1a, Extended Data Fig. 3c–f). We therefore conclude that UDP–glucose and isopentenyl pyrophosphate are not likely to be physiological substrates for mammalian viperins; although it remains a possibility that other eukaryotes use viperin homologues to perform distinct functions. Analogous competition experiments demonstrated that deoxyCTP is also not a substrate of rVIP (Extended Data Fig. 3g).

Incubation of rVIP with SAM and CTP deuterated at the 2′, 3′, 4′, 5′ and 5 positions (deuCTP), increased the $-m/z$ of 5′-dA from 250.1 to 251.1, consistent with the transfer of one deuterium from deuCTP to 5′-dA. Additionally, ddhCTP from this reaction exhibited a $-m/z$ of 468.1, indicating that the deuterium abstracted by 5′-dA did not return to the product (Extended Data Fig. 3h, i). Derivatives

[1]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY, USA. [2]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, USA. [3]Department of Microbiology and Immunology, Albert Einstein College of Medicine, Bronx, NY, USA. [4]Institute for Protein Innovation, Boston, MA, USA. [5]These authors contributed equally: Anthony S. Gizzi, Tyler L. Grove. *e-mail: tyler.grove@einstein.yu.edu; steve.almo@einstein.yu.edu

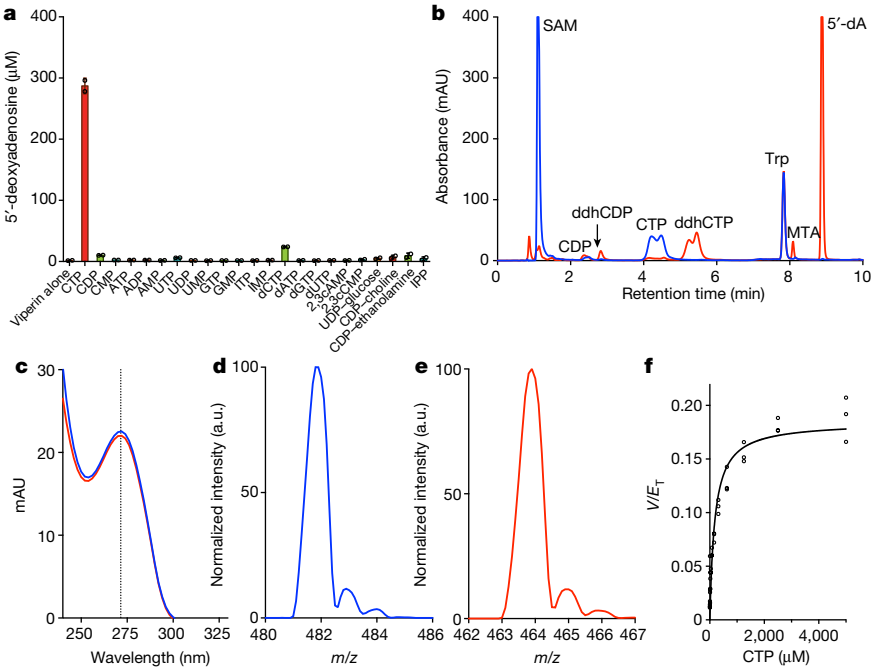


Fig. 1 | Substrate specificity of viperin. a, A panel of nucleotides was mixed with rVIP and SAM, and the resulting 5′-dA measured. Data are from two independent experiments. **b**, High-performance liquid chromatography analysis showing viperin-mediated conversion of CTP to a new product (time = 0, blue; time = 45 min, red). **c**, UV-visible spectrum of CTP (blue) and the new product (ddhCTP, red). Absorbance maximum at 271 nm (dotted line). **d**, **e**, Mass spectrometry for CTP (**d**, $-m/z = 482.1$) and ddhCTP (**e**, $-m/z = 464.1$). All results have been repeated at least three times. **f**, Kinetic analysis of rVIP with CTP: K_m for CTP = 182.8 ± 27.6 μM and $V_{max} = 0.185 \pm 0.007$ min^{-1} (V/E_T, initial velocity of the reaction divided by the total enzyme concentration). Data are mean ± s.d. from three independent experiments.

of site-specifically deuterated CTP demonstrated that 5′-dA• initiates the reaction by uniquely abstracting the hydrogen from the 4′ position of CTP (see Supplementary Information for details, Extended Data Figs. 2, 4). On the basis of this observation, a provisional mechanism is outlined in Fig. 2d, in which viperin uses the 5′-dA• to abstract the hydrogen atom at the 4′ position of the ribose of CTP, subsequently allowing for the loss of the 3′-hydroxyl group with general acid assistance. The resulting resonance-stabilized radical cation is then reduced by one electron to yield the designated product. This mechanism has precedent from model studies of the radiolytic cleavage of single-stranded DNA, wherein generation of a 4′-deoxyribosyl radical causes heterolytic dissociation of the 3′ phosphate group, resulting in a 3′-cation-4′-yl radical[11]. The source of the additional electron needed to reduce intermediate three (or four) in the viperin-catalysed reaction is currently unclear. However, we propose that—similar to other radical-SAM enzyme reactions[12]—the electron probably derives from a reduced 4Fe–4S cluster, which suggests that viperin requires two electrons to complete each turnover: one to generate the 5′-dA• and another to reduce intermediate three (or four).

CMPK2 is always immediately adjacent to viperin in vertebrate genomes and is co-transcribed with viperin during IFN stimulation[13], suggesting a functional linkage. Human CMPK2 (hCMPK2) was previously reported to catalyse the ATP-dependent phosphorylation of the monophosphates CMP, UMP and dCMP to the corresponding diphosphates[14]. By contrast, we find that hCMPK2 exhibits strong preference for CDP and UDP as substrates, yielding CTP and UTP, respectively (Extended Data Fig. 5a). Notably, when provided with ddhCDP, hCMPK2 displayed a tenfold-lower activity for producing ddhCTP when compared to the rate of CTP and UTP formation (Extended Data Fig. 5a, Supplementary Table 2). Therefore, on the basis of the synteny and coordinated expression of viperin and CMPK2, and the available biochemical data, we propose that CMPK2 primarily functions to ensure sufficient substrate (that is, CTP produced from CDP, or by

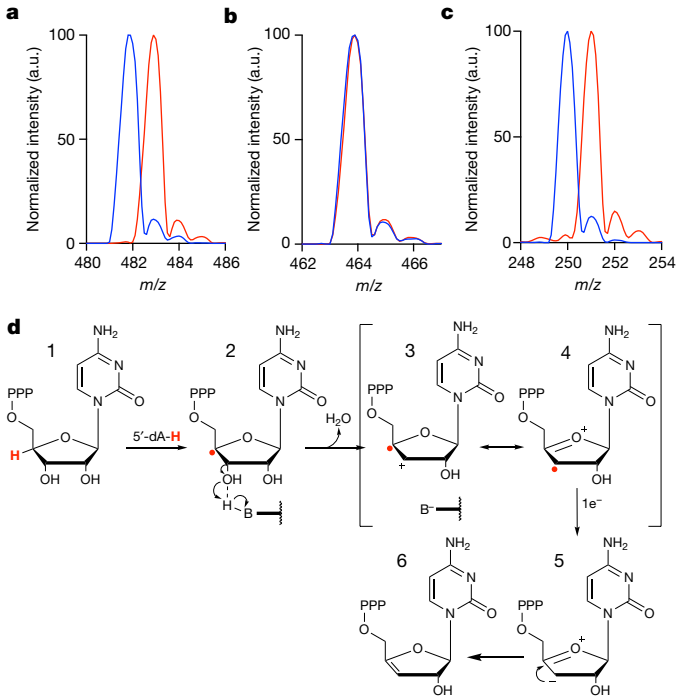


Fig. 2 | Proposed mechanism for formation of ddhCTP. a, m/z of CTP (blue) or 4′-^{2}H -CTP (red). **b**, Mass spectrometry of ddhCTP from reactions with either CTP (blue), or 4′-^{2}H -CTP (red) and rVIP. Deuterium from 4′-^{2}H -CTP is not retained in ddhCTP as products have the same $-m/z$ 464.1. **c**, 5′-dA derived from 4′-^{2}H -CTP (red trace) increases by one mass unit owing to the incorporation of deuterium. These experiments have been repeated at least three times with similar results. **d**, After hydrogen atom abstraction at the 4′ position of CTP, general base-assisted loss of the 3′ hydroxyl group leads to a carbocation/radical intermediate that is reduced by one electron to yield the ddhCTP product.

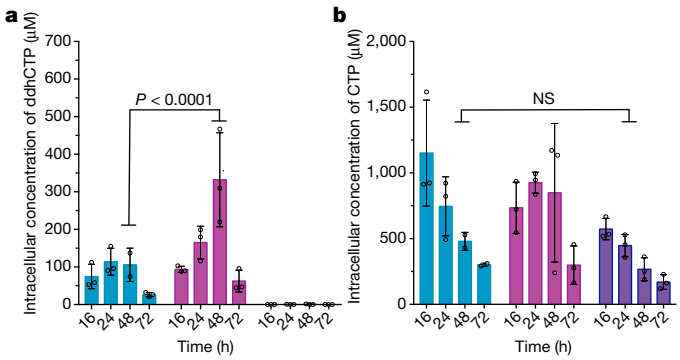


Fig. 3 | Expression of viperin in HEK293 cells produces ddhCTP. **a**, Cells expressing hVIP (aqua), hVIP and hCMPK2 (pink) or empty vector (purple). Analysis of ddhCTP formation indicates that the cells with hVIP and hCMPK2 cells show a statistically significant increase in ddhCTP formation over viperin alone 48 h after transfection. In cells with empty vector, ddhCTP levels were undetectable. **b**, Intracellular concentrations of CTP did not differ significantly over time. Data are mean ± s.d. from three biologically independent samples; *P* value is from a two-way ANOVA with Tukey's post hoc test. NS, not significant.

CTP synthetase acting on UTP) for viperin-mediated production of ddhCTP during viral infection.

To demonstrate that ddhCTP can be produced in mammalian cytosol, we generated a series of human viperin (hVIP) (83% identical to rVIP) and hCMPK2 expression constructs for transient transfection in HEK293T cells (Supplemenary Table 3, Extended Data Fig. 5b). As HEK293T cells do not express viperin in the presence or absence of IFN[15], ddhCTP production would not be expected in the absence of exogenous viperin expression. HEK293T cells were transfected with a control plasmid, hVIP alone, hCMPK2 alone or both hVIP and hCMPK2, and collected at defined times for LC–MS analysis (see Supplementary Information for details). In all cases, over a 72-h time course, the overall nucleotide pool consistently decreases, probably owing to limiting nutrient levels, though the overall growth and cell viability are not affected (Supplementary Table 4). In addition, at each time point there are no statistically significant differences in total nucleotide concentrations between the control, hVIP, and hVIP and hCMPK2 treated cells (Extended Data Fig. 6a–d). Notably, HEK293T cells transfected with control plasmid exhibited ddhCTP levels below our limit of detection of approximately 400 fmol (Fig. 3a, right), whereas HEK293T cells transfected with the hVIP plasmid exhibited high intracellular ddhCTP levels (approximately 75 μM at 16 h post-transfection) (Fig. 3a, left), which decreases to approximately 35 μM after 72 h. Cotransfection with hVIP and hCMPK2 plasmids resulted in an approximately fourfold increase in the amount of ddhCTP to approximately 330 μM at 48 h (Fig. 3a, middle, purple, $P < 0.0001$). Moreover, coexpression of hVIP and hCMPK2 causes the ratio of ddhCTP to CTP concentrations to increase over time, whereas overexpression of hVIP alone results in a constant ratio of ddhCTP to CTP (Extended Data Fig. 6e). This behaviour may enable viperin to continue generating ddhCTP even though roughly 30% of the total cellular pool of cytidine triphosphates is present as ddhCTP at 48 h. These observations demonstrate that in vivo viperin is essential for production of ddhCTP, and suggest that CMPK2 may function to ensure that CTP is not limiting in the presence of viperin.

It is well documented that viperin expression can be robustly induced in immune cells by interferon, lipopolysaccharide and double-stranded RNA analogues[16]. Therefore, we cultured immortalized murine macrophages (RAW264.7) in the presence or absence of IFNα in serum-free medium, as it has previously been shown that RAW264.7 cells express viperin in an IFNα-sensitive fashion[2]. When collected after 19 h, the concentration of ddhCTP was shown to be highly dependent on the concentration of IFNα (Extended Data Fig. 7a). RAW264.7 cells cultured in the presence of 250 ng ml^{-1} IFNα generated intracellular concentrations of ddhCTP reaching nearly 350 μM, whereas the intracellular concentrations of ATP, UTP and CTP were unaltered (Extended Data Fig. 7b). Analogous to the behaviour observed in HEK293T cells cotransfected with hVIP and hCMPK2, treatment of RAW264.7 cells with 250 ng ml^{-1} IFNα resulted in ddhCTP representing a sizable proportion (that is, 30%) of the total cytidine triphosphate pool (ddhCTP (approximately 350 μM) to CTP (approximately 800 μM)), whereas the level of CTP remained unchanged. This behaviour suggests that the viperin-mediated inhibition of viral replication is not a consequence of the limitation of the available pool of intracellular CTP, or other nucleotides, but rather is dependent on the generation of relevant concentrations of ddhCTP.

Because members of the flavivirus family are known to be sensitive to the catalytic activity of viperin[17], and because of the resemblance of ddhCTP to known polymerase chain terminators, we examined the effect of ddhCTP on dengue virus RNA-dependent RNA polymerase (RdRp) activity. First, we demonstrated that ddhCTP is a substrate for dengue virus RdRp using a primed-template assay[18] (Extended Data Fig. 8a, b). Addition of CTP, 3′-dCTP or ddhCTP led to incorporation of all of these nucleotides (Extended Data Fig. 8b). As expected, addition of UTP to the CMP-incorporated RNA led to further extension to the end of template (Extended Data Fig. 8b, c). However, addition of UTP to the 3′-dCMP- or ddhCMP-incorporated RNA did not support robust extension (Extended Data Fig. 8b, c), as expected for the action of chain terminators. A more stringent test of the effectiveness of a chain terminator is direct competition with natural ribonucleotides. Therefore, we evaluated RNA synthesis in the presence of increasing concentrations of ddhCTP or 3′-dCTP (Fig. 4a). Both ddhCTP and 3′-dCTP were incorporated and inhibited production of full-length RNA (Fig. 4a). Additionally, by titrating ddhCTP at different concentrations of CTP we determined the relative efficiency of the use of ddhCTP compared to CTP for dengue virus RdRp, as well as the RdRp from West Nile Virus (WNV), another pathogenic flavivirus (Fig. 4b, Extended Data Fig. 8e–j). This analysis yielded a 135- and 59-fold difference in the use of ddhCTP relative to CTP for dengue virus and WNV RdRps, respectively. We also evaluated two additional members of the flavivirus family, Zika virus (ZIKV) and hepatitis C virus RdRps. Both of these RdRps were susceptible to inhibition by ddhCTP utilization and chain termination (Fig. 4c, d), consistent with studies demonstrating the antiviral activity of viperin against these viruses[19–21]. These data suggest that the flavivirus RdRps would be susceptible to inhibition by ddhCTP during replication (Extended Data Fig. 9g). Given the efficiency of use and genome size, it is calculated that even a probability of incorporating the ddhCTP chain terminator during replication of approximately 1% would result in considerable reduction of full-length genomes (Extended Data Fig. 9g). To determine whether our observations with the flavivirus RdRps extend to RdRps from other supergroups, we evaluated members of supergroup I. Specifically, we used the RdRps from human rhinovirus type C (HRV-C) and poliovirus, which are both members of the picornavirus family (Extended Data Fig. 9). Direct-incorporation experiments revealed the use of both ddhCTP and 3′-dCTP by HRV-C RdRp (Extended Data Fig. 9b). However, in the presence of other rNTPs, both HRV-C and poliovirus RdRp were poorly inhibited by ddhCTP (Extended Data Fig. 9c–f), even though both are inhibited by 3′-dCTP. On the basis of this data, we conclude that not all RdRps are sensitive to ddhCTP, suggesting that different viruses will exhibit a range of susceptibilities to viperin expression in vivo.

The in vitro enzymatic characterizations suggest that ddhCTP would be sufficient for the in vivo inhibition of viral replication. First, we demonstrated that synthetic ddhC nucleoside was capable of traversing the plasma membrane of Vero and HEK293T cells and being metabolized to ddhCTP (1 mM synthetic ddhC resulted in the intracellular accumulation of 129 μM and 78 μM ddh-CTP after 24 h, respectively) (Extended Data Fig. 5c, d). Next, we used the historical African strain MR766 (Uganda 1947)[22] and two contemporary strains, PRVABC59 (Puerto Rico; 2015)[23] and R103451 (Honduras; 2016, GenBank:

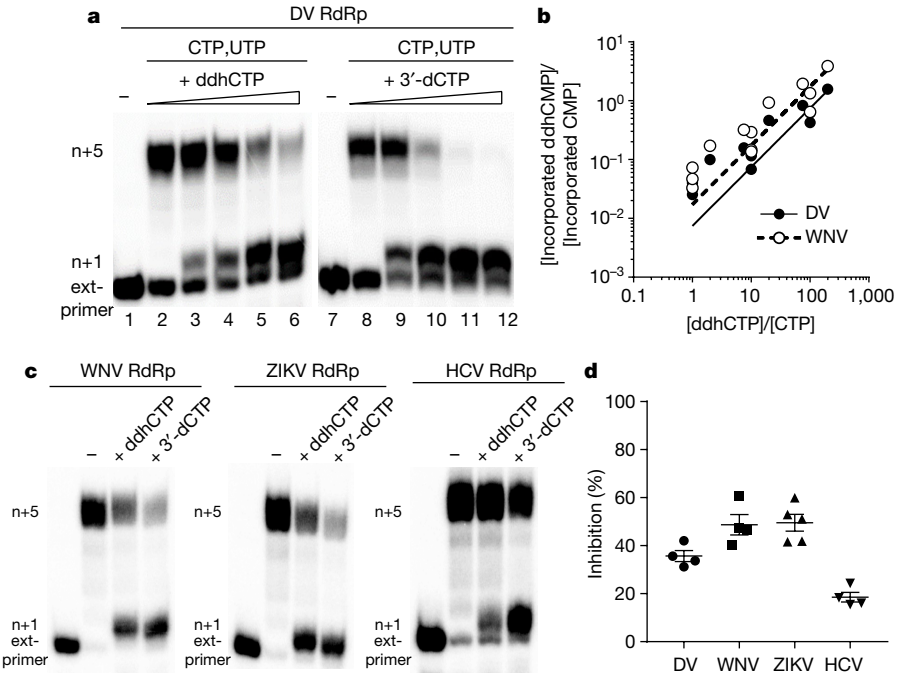


Fig. 4 | ddhCTP inhibits flavivirus RdRps by a chain termination mechanism. a, Primer extension assays show that ddhCTP (0, 1, 10, 100 and 300 μM) inhibits dengue virus (DV) RdRp. A trinucleotide is used to prime RNA synthesis in the presence of nucleotides (α^{32}P-NTPs) producing a 15-nt product that we refer to as extended primer (ext-primer). Additional incorporation of nucleotides leads to production of n + 1 and/or n + 5 product. Experiments repeated independently four times with similar results. **b**, Plot of [incorporated ddhCMP]/[incorporated CMP] versus [ddhCTP]/[CTP]. Data are fit to lines with slopes of 0.0074 ± 0.0006 (DV) and 0.017 ± 0.002 (WNV). At each ratio of [ddhCTP]/[CTP] a total of at least three independent experiments were preformed, with a total sample size of 24. **c**, Inhibition of the RdRp from WNV, ZIKV and hepatitis C virus (HCV) by either ddhCTP or 3′-dCTP (100 μM). Experiments repeated independently four to five times with similar results. **d**, Percentage of inhibition shown. Data are mean ± s.e.m. from four independent experiments. Gel shown in panel **c** was quantified by using ImageQuant TL software (GE).

KX262887), to evaluate the antiviral activity of ddhC towards ZIKV replication and release from Vero cells. Treatment with ddhC resulted in a reduction in ZIKV titres of one to two orders of magnitude, which was dependent on dose, multiplicity of infection (MOI), duration of infection and strain (Supplementary Table 5). For example, at a MOI of 0.1, we observed 50–200-fold reduction in viral titre for ZIKV MR766 at all time points (Fig. 5a), with reductions of 5–50-fold also observed at a MOI of 1.0 (Extended Data Fig. 10a). ZIKV R103451 (Honduras) and ZIKV PRVABC59 (Puerto Rico) exhibited analogous sensitivities to ddhC treatment (Supplementary Table 5, Extended Data Fig. 10a–c). The reduction in viral release was not a result of ddhC cytotoxicity, as incubation with 1 mM ddhC did not alter Vero cell viability (Fig. 5b, see Supplementary Information for details). These results, taken together with the in vitro enzymatic analyses, are consistent with a model in which ddhC-derived ddhCTP inhibits viral replication through premature chain termination of RdRp products.

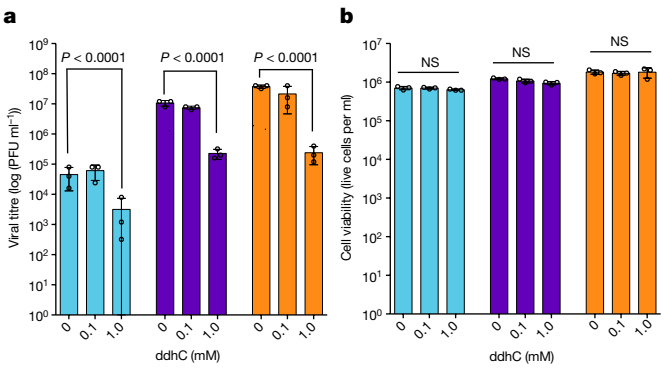


Fig. 5 | ddhC reduces ZIKV release in Vero cells. a, Vero cells were treated with increasing concentrations of ddhC for 24 h and infected with ZIKV MR766 (Uganda) at an MOI of 0.1. PFU, plaque-forming unit. **b**, For viability studies, Vero cells were treated with increasing concentrations of ddhC under the same culture conditions used for the antiviral experiment described in **a**. Data are mean ± s.d. from three biologically independent samples; *P* values are from a two-way ANOVA with Dunnett's post hoc analysis.

Of the hundreds of genes stimulated by IFN, most appear to function as negative effectors of viral activity, though their mechanisms of action remain to be defined. Here we propose a new paradigm for the antiviral function of viperin, which relies on its intrinsic catalytic activity to generate ddhCTP, a previously undescribed replication-chain terminator. To our knowledge, viperin is the only human protein that produces a small molecule capable of directly inhibiting viral replication machinery. Importantly, overexpression of viperin and production of ddhCTP does not appear to adversely affect the growth rate or viability of HEK293T or Vero cells. This observation indicates that the host RNA and DNA polymerases are not negatively affected by ddhCTP and have developed protective mechanisms to exclude incorporation or excise this compound during nucleic-acid synthesis; mechanistic studies on the use of ddhCTP by host polymerases will be an important area for future investigation. In addition to its inhibitory effect on viral RdRp activity, it is possible that viperin possesses additional antiviral functions. For example, despite reports that HRV infection induces viperin expression, ddhCTP does not appear to act as an effective chain terminator for the HRV RdRp[24]. Furthermore, in the case of human cytomegalovirus, viperin expression results in enhanced infectivity, possibly through alterations in cellular metabolism and disruption of the actin cytoskeleton[25]. It is probable that different pathogens are responsive to distinct subsets of the IFN-inducible genes and—given the range of modulatory effects it has on viruses—that viperin synergizes with other host- and pathogen-encoded genes.

Data availability

All data supporting the conclusions of the paper are available on reasonable request from the authors.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0238-4.

Received: 17 July 2017; Accepted: 27 April 2018;
Published online: 20 June 2018

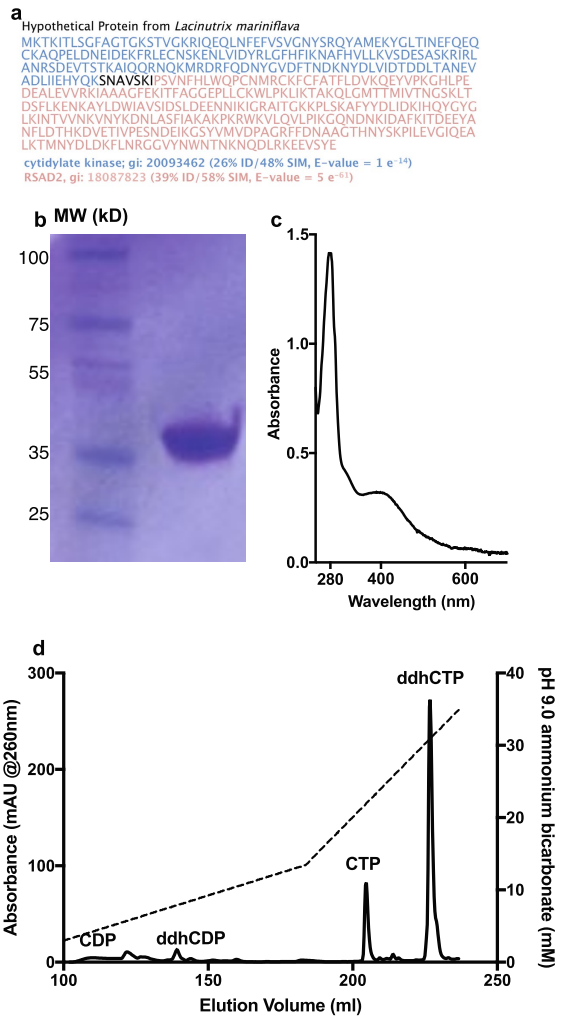
1. Molinari, N. A. M. et al. The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine* **25**, 5086–5096 (2007).
2. Tang, H. B. et al. Viperin inhibits rabies virus replication via reduced cholesterol and sphingomyelin and is regulated upstream by TLR4. *Sci. Rep.* **6**, 30529 (2016).
3. Helbig, K. J. & Beard, M. R. The role of viperin in the innate antiviral response. *J. Mol. Biol.* **426**, 1210–1219 (2014).
4. Seo, J. Y., Yaneva, R. & Cresswell, P. Viperin: a multifunctional, interferon-inducible protein that regulates virus replication. *Cell Host Microbe* **10**, 534–539 (2011).
5. Hover, B. M., Loksztejn, A., Ribeiro, A. A. & Yokoyama, K. Identification of a cyclic nucleotide as a cryptic intermediate in molybdenum cofactor biosynthesis. *J. Am. Chem. Soc.* **135**, 7019–7032 (2013).
6. Fenwick, M. K., Li, Y., Cresswell, P., Modis, Y. & Ealick, S. E. Structural studies of viperin, an antiviral radical SAM enzyme. *Proc. Natl Acad. Sci. USA* **114**, 6806–6811 (2017).
7. Kennedy, A. D. et al. Complete nucleotide sequence analysis of plasmids in strains of *Staphylococcus aureus* clone USA300 reveals a high level of identity among isolates with closely related core genome sequences. *J. Clin. Microbiol.* **48**, 4504–4511 (2010).
8. Yokoyama, K., Numakura, M., Kudo, F., Ohmori, D. & Eguchi, T. Characterization and mechanistic study of a radical SAM dehydrogenase in the biosynthesis of butirosin. *J. Am. Chem. Soc.* **129**, 15147–15155 (2007).
9. Honarmand Ebrahimi, K. et al. The radical-SAM enzyme Viperin catalyzes reductive addition of a 5′-deoxyadenosyl radical to UDP-glucose *in vitro*. *FEBS Lett.* **591**, 2394–2405 (2017).
10. Lee, H. *A Proposed Mechanism for the Radical SAM Enzyme Viperin*. BSc thesis, Univ. of Illinois (2017).
11. Giese, B., Beyrich-Graf, X., Erdmann, P., Petretta, M. & Schwitter, U. The chemistry of single-stranded 4′-DNA radicals: influence of the radical precursor on anaerobic and aerobic strand cleavage. *Chem. Biol.* **2**, 367–375 (1995).
12. Grove, T. L. et al. A substrate radical intermediate in catalysis by the antibiotic resistance protein Cfr. *Nat. Chem. Biol.* **9**, 422–427 (2013).
13. Kambara, H. et al. Negative regulation of the interferon response by an interferon-induced long non-coding RNA. *Nucleic Acids Res.* **42**, 10668–10680 (2014).
14. Xu, Y., Johansson, M. & Karlsson, A. Human UMP-CMP kinase 2, a novel nucleoside monophosphate kinase localized in mitochondria. *J. Biol. Chem.* **283**, 1563–1571 (2008).
15. Teng, T. S. et al. Viperin restricts chikungunya virus replication and pathology. *J. Clin. Invest.* **122**, 4447–4460 (2012).
16. Wang, B. et al. Viperin is induced following toll-like receptor 3 (TLR3) ligation and has a virus-responsive function in human trophoblast cells. *Placenta* **36**, 667–673 (2015).
17. Jiang, D. et al. Identification of five interferon-induced cellular proteins that inhibit West Nile virus and dengue virus infections. *J. Virol.* **84**, 8332–8341 (2010).
18. Van Slyke, G. A. et al. Sequence-specific fidelity alterations associated with West Nile virus attenuation in mosquitoes. *PLoS Pathog.* **11**, e1005009 (2015).
19. Panayiotou, C. et al. Viperin restricts Zika virus and tick-borne encephalitis virus replication by targeting NS3 for proteasomal degradation. *J. Virol.* **92**, e02054-17 (2018).
20. Szretter, K. J. et al. The interferon-inducible gene viperin restricts West Nile virus pathogenesis. *J. Virol.* **85**, 11557–11566 (2011).
21. Wang, S. et al. Viperin inhibits hepatitis C virus replication by interfering with binding of NS5A to host protein hVAP-33. *J. Gen. Virol.* **93**, 83–92 (2012).
22. Dick, G. W., Kitchen, S. F. & Haddow, A. J. Zika virus. I. Isolations and serological specificity. *Trans. R. Soc. Trop. Med. Hyg.* **46**, 509–520 (1952).
23. Lanciotti, R. S., Lambert, A. J., Holodniy, M., Saavedra, S. & Signor, L. d. C C. Phylogeny of Zika Virus in Western hemisphere, 2015. *Emerg. Infect. Dis.* **22**, 933–935 (2016).
24. Proud, D. et al. Gene expression profiles during *in vivo* human rhinovirus infection: insights into the host response. *Am. J. Respir. Crit. Care Med.* **178**, 962–968 (2008).
25. Seo, J. Y. & Cresswell, P. Viperin regulates cellular lipid metabolism during human cytomegalovirus infection. *PLoS Pathog.* **9**, e1003497 (2013).

Acknowledgements We thank S. J. Booker for helpful discussions, L. Nordstroem (Chemical Synthesis & Biology Core Facility) for synthesis of ddhC and R. Sharma and J. Perryman for assistance with construction of RdRp expression plasmids and purification of RdRp enzymes. This work was supported by National Institutes of Health (NIH) Grants R21-AI133329 (T.L.G. and S.C.A.), P01-GM118303-01 (J. A. Gerlt and S.C.A.), U54-GM093342 (J. A. Gerlt and S.C.A.), U54-GM094662 (S.C.A.), R01-AI045818 (C.E.C.), Pennsylvania State University Start-Up Funds (J.J.), and the Price Family Foundation (S.C.A.). We acknowledge the Albert Einstein Anaerobic Structural and Functional Genomics Resource (http://www.nysgxrc.org/psi3/anaerobic.html).

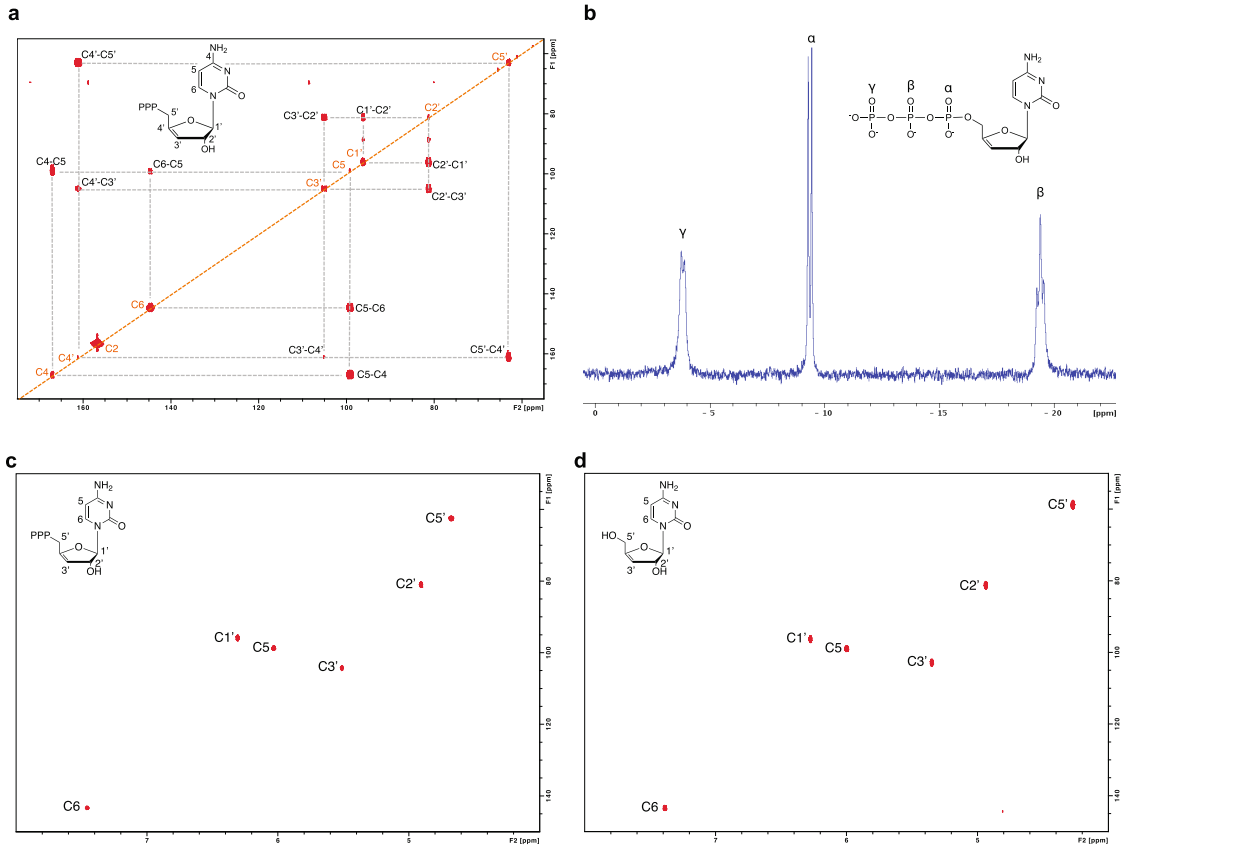
Author contributions A.S.G., T.L.G., J.J.A., C.E.C. and S.C.A. designed the research; A.S.G. and T.L.G. contributed equally; A.S.G. and T.L.G. prepared protein and performed experiments; J.J.A. performed polymerase biochemistry; J.J. performed ZIKV release assays; Q.D. prepared isotopologues; R.K.J. and K.C. provided advice on virologic experiments and performed statistical analysis; S.M.C. performed NMR measurements; S.J.G. prepared HEK293T cells; N.G.D. prepared RAW264.7 cells; all authors analysed data. T.L.G., J.D.L., A.R.B., C.E.C. and S.C.A. supervised research. A.S.G., T.L.G., J.J.A., C.E.C. and S.C.A. wrote the manuscript.

Competing interests A.S.G., T.L.G., J.J.A., C.E.C. and S.C.A. are co-inventors on a U.S. provisional patent application (No. 62/548,425; filed by S.C.A.) that incorporates discoveries described in this manuscript.

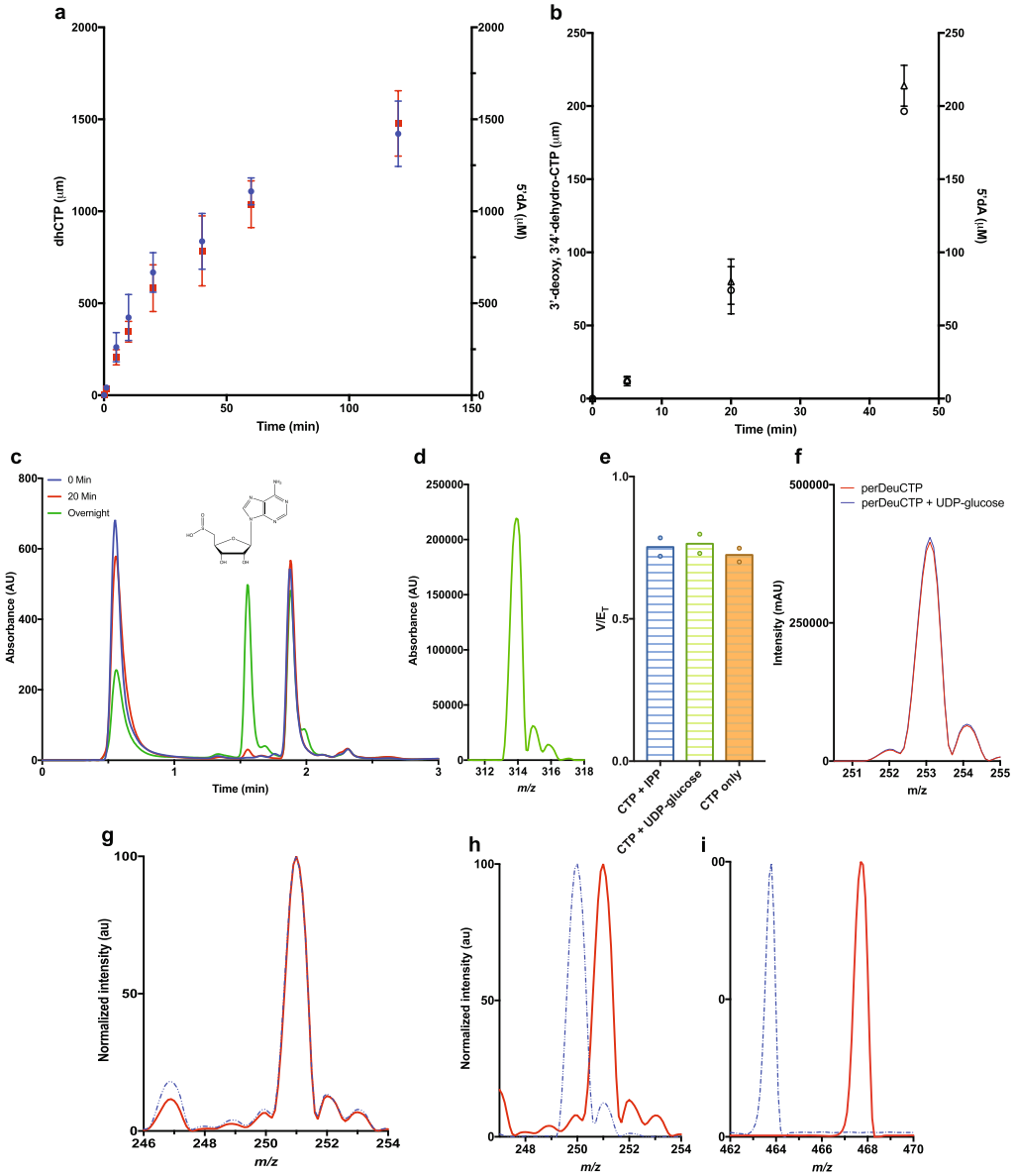
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0238-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0238-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.L.G. or S.C.A.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



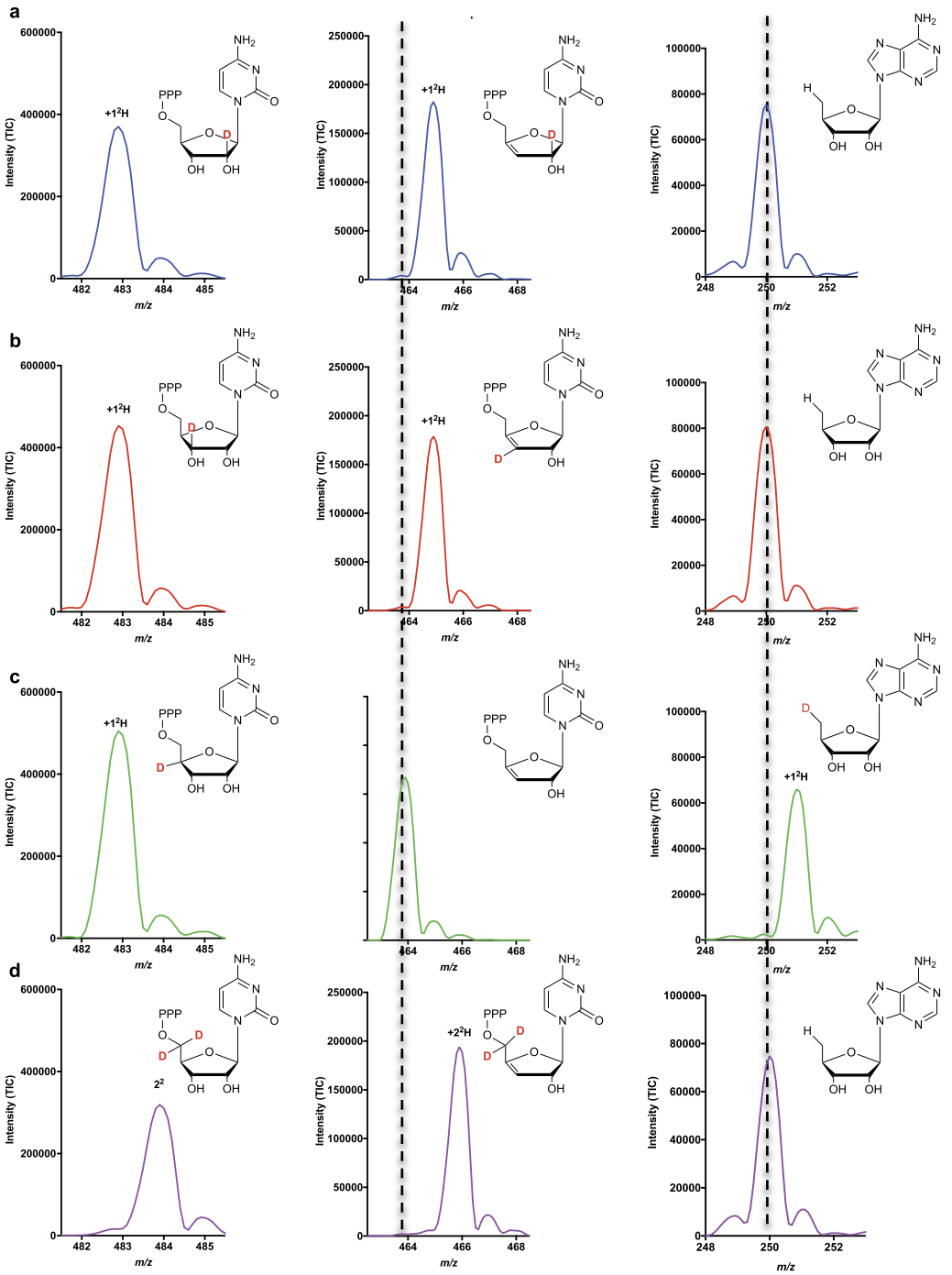
Extended Data Fig. 1 | Purification of rVIP and ddhCTP. a, Amino acid sequence from a *Lacinutrix mariniflava* fusion gene product of CMPK2- and a viperin-like protein. **b**, SDS–PAGE analysis after affinity and size-exclusion chromatography. The protein corresponding to amino acid residues 51–361 has a predicted molecular mass of 38.36 kDa. This construct was chosen because approximately 100 mg of protein could be purified from a 2-l fermentation. In addition, the protein is soluble at concentrations greater than 2 mM. **c**, UV-visible spectrum of purified rVIP (29.5 μM, UV 280/400 ratio of 4.2). **d**, Purification of ddhCTP using ammonium bicarbonate pH 9, with an elution gradient (dashed line) from 0.2 M to 0.8 M over 200 column volumes. All results have been repeated at least three times.



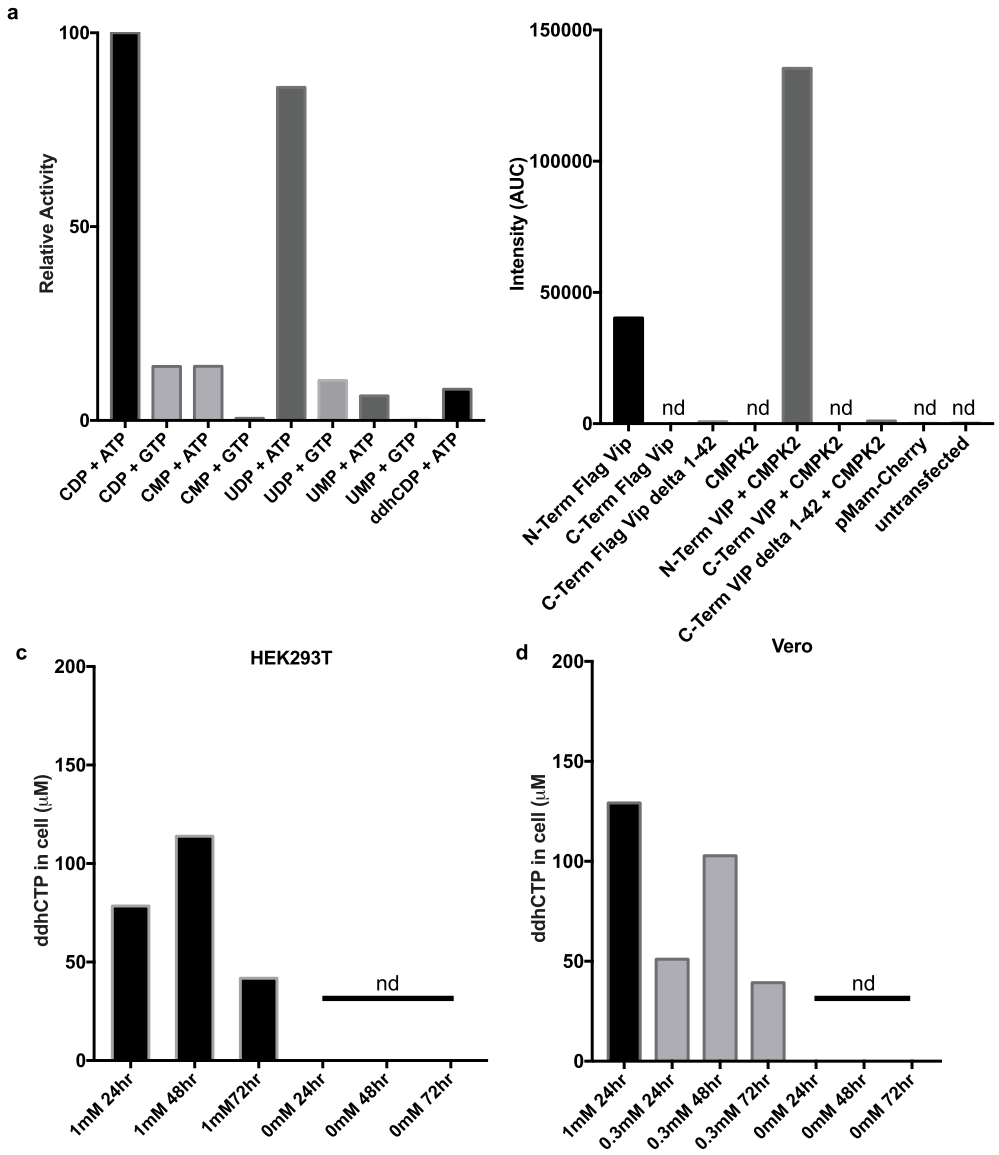
Extended Data Fig. 2 | NMR spectroscopy of ddhCTP. a, ^{13}C–^{13}C COSY spectrum of $^{13}C_9{}^{15}N_3$–ddhCTP. The assignments for the observed correlations of the ^{13}C-connectivities are indicated with the grey dotted lines. **b**, ^{31}P NMR spectra (300 MHz) of ddhCTP (1 mM) in D_2O at 300 K. Three resonance peaks at –19.5 (triplet), –9.5 (doublet) and –3.9 (doublet) p.p.m. correspond to the beta, alpha and gamma phosphates of ddhCTP, respectively. **c**, 2D-HSQC NMR spectra collected on purified 1 mM ddhCTP in D_2O. **d**, 2D-HSQC NMR spectra collected on 1 mM synthetic ddhC in D_2O. All experiments have been repeated twice.



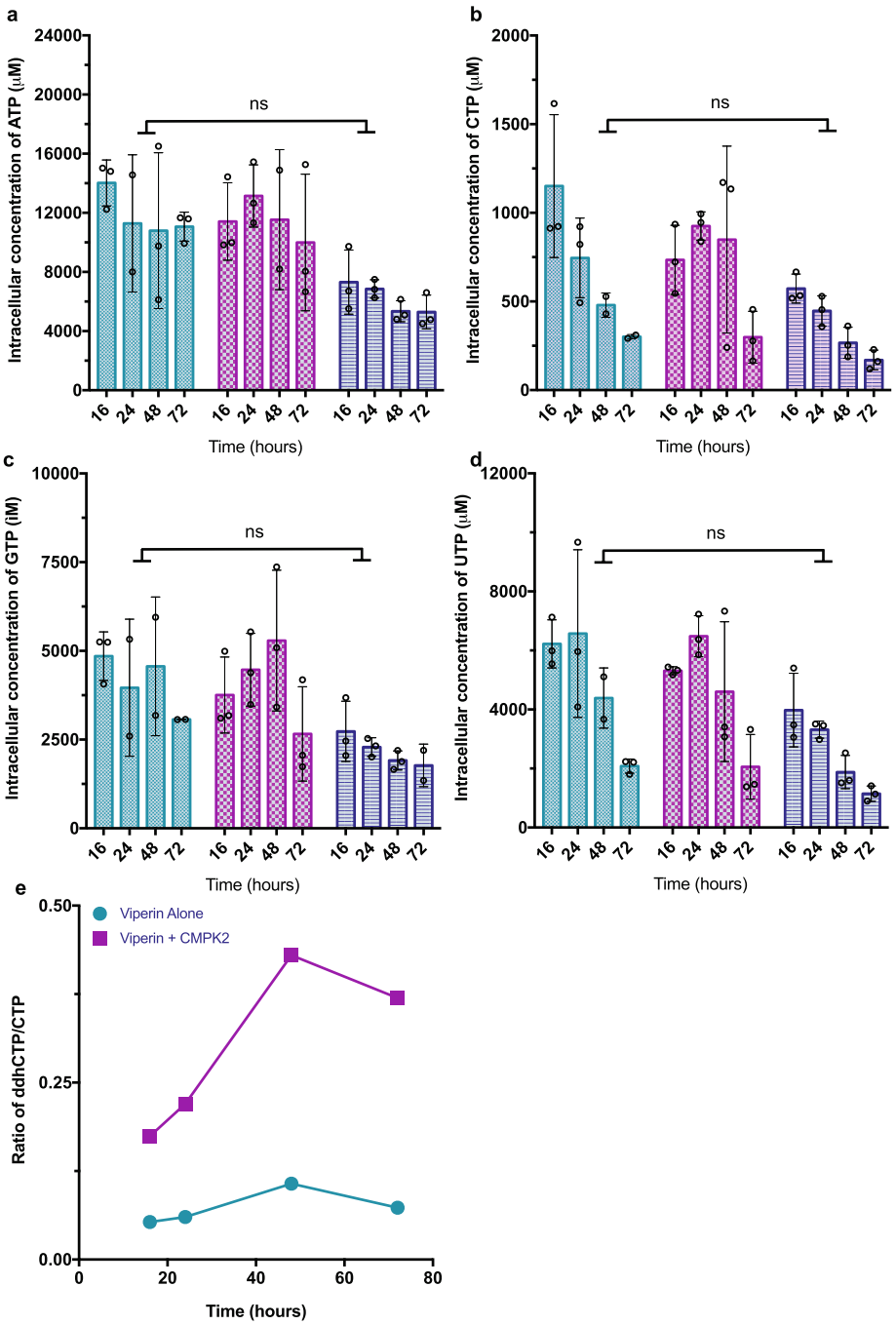
Extended Data Fig. 3 | rVIP produces a 1:1 stoichiometry of 5′-dA and ddhCTP and reacts specifically with CTP. a, Formation of ddhCTP (red squares) and 5′-dA (blue circles) from CTP and SAM in the presence of dithionite and 100 μM rVIP. ddhCTP is formed at roughly stoichiometric amounts with that of 5′-dA. Data are mean ± s.d. from three replicates. **b**, Formation of ddhCTP (open triangle,) and 5′-dA (open circle) from CTP and SAM in the presence of the flavodoxin, flavodoxin reductase and NADPH using 100 μM rVIP. Data are mean ± s.d. from three replicates. ddhCTP and 5′-dA are formed at roughly stoichiometric concentrations. The production of ddhCTP by this enzyme-driven reducing system indicates that ddhCTP formation is not the consequence of a side reaction with dithionite. **c**, High-performance liquid chromatography analysis (0 min, blue trace; 20 min, red trace; 12 h, green trace) showing the generation of a new peak at 1.55 min in the 12 h sample corresponding to a 5′-dA–dithionite adduct in the presence of 100 μM rVIP, 1 mM SAM and 10 mM IPP. **d**, Corresponding mass spectra in ESI negative mode of the peak occurring at 1.55 min in the 12 h sample. The 5′-dA–dithionite conjugate was calculated to have an exact mass of 315 Da and an *m*/*z* of 314.1. These results have been repeated twice. **e**, The rate of 5′-dA formed by 100 μM rVIP in the presence of 1 mM SAM and 1 mM CTP alone or 1 mM CTP with 10 mM IPP or UDP–glucose. Data are mean ± s.d. from three independent experiments. **f**, Mass spectrum traces of 5′-dA by ESI+. Reactions were conducted with 100 μM rVIP, 1 mM SAM and 1mM deuCTP with or without 10 mM UDP–glucose. The mass spectrum of 5′-dA produced during these reactions shows only the presence of deuterium, which derives from deuCTP, even when UDP–glucose is present at a tenfold higher concentration. An *m*/*z* of 252.1 represents the natural abundance peak of 5′-dA, an *m*/*z* of 253.1 indicating the addition of one deuterium. **g**, Mass spectrum trace showing –*m*/*z* of 5′-dA formed by combining 100 μM rVIP with 1 mM deuCTP (dotted blue trace) or 1 mM deuCTP with 1 mM deoxyCTP (red trace). The *y*-axis of each spectrum was normalized to 100% with arbitrary units (au) to enable direct comparison between each sample. The 5′-dA produced during this reaction has an *m*/*z* of 251.1, which is only consistent with rVIP abstracting a deuteron from deuCTP and not acting on the deoxyCTP (that is, lack of *m*/*z* 250.1). **h**, **i**, Mass spectrum trace showing –*m*/*z* of 5′-dA (**h**) or the new product (**i**), formed by combining 100 μM rVIP with either 1 mM CTP (dotted blue trace) or 1 mM deuCTP (red trace). When rVIP was incubated with SAM and CTP deuterated at the 2′, 3′, 4′, 5′ and 5 positions (deuCTP), the −*m*/*z* of 5′-dA increased from 250.1 to 251.1, consistent with the transfer of one deuterium from deuCTP to 5′-dA●. When ddhCTP from the reaction was analysed by mass spectormetry, the product exhibited a −*m*/*z* of 468.1, indicating that the deuterium abstracted by 5′-dA during catalysis did not return to the product. The *y*-axis of each spectrum was normalized to 100% with arbitrary units (au) to enable direct comparison between each sample. These results have been repeated at least twice.



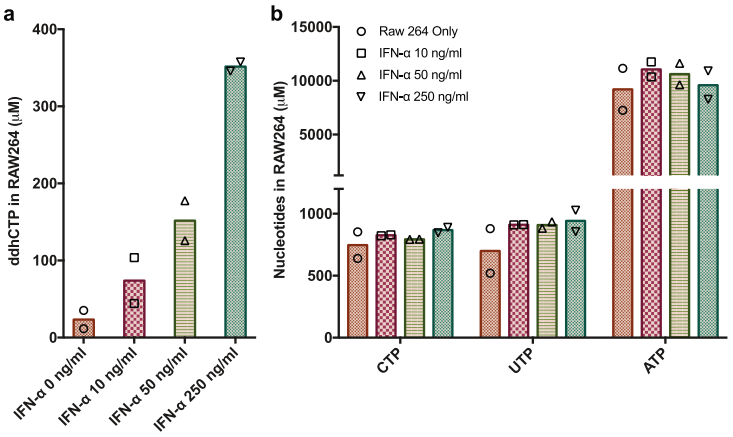
Extended Data Fig. 4 | Viperin abstracts the 4′-H from CTP. a–d, Using CTP with deuterium (^{2}H denoted with a red D) incorporated at either the 2′-^{2}H (**a**), 3′-^{2}H (**b**), 4′-^{2}H (**c**) or 5′-^{2}H$_2$ (**d**) (left column), we were able to monitor the loss of deuterium from the resulting product (middle column) and gain of a deuterium in the resulting 5′-dA (right column). The 5′-dA –*m*/*z* increases by one only in reactions containing CTP with a 4′-^{2}H (**c**, right column). Natural abundance peaks are denoted with dashed vertical lines. All experiments were repeated twice.



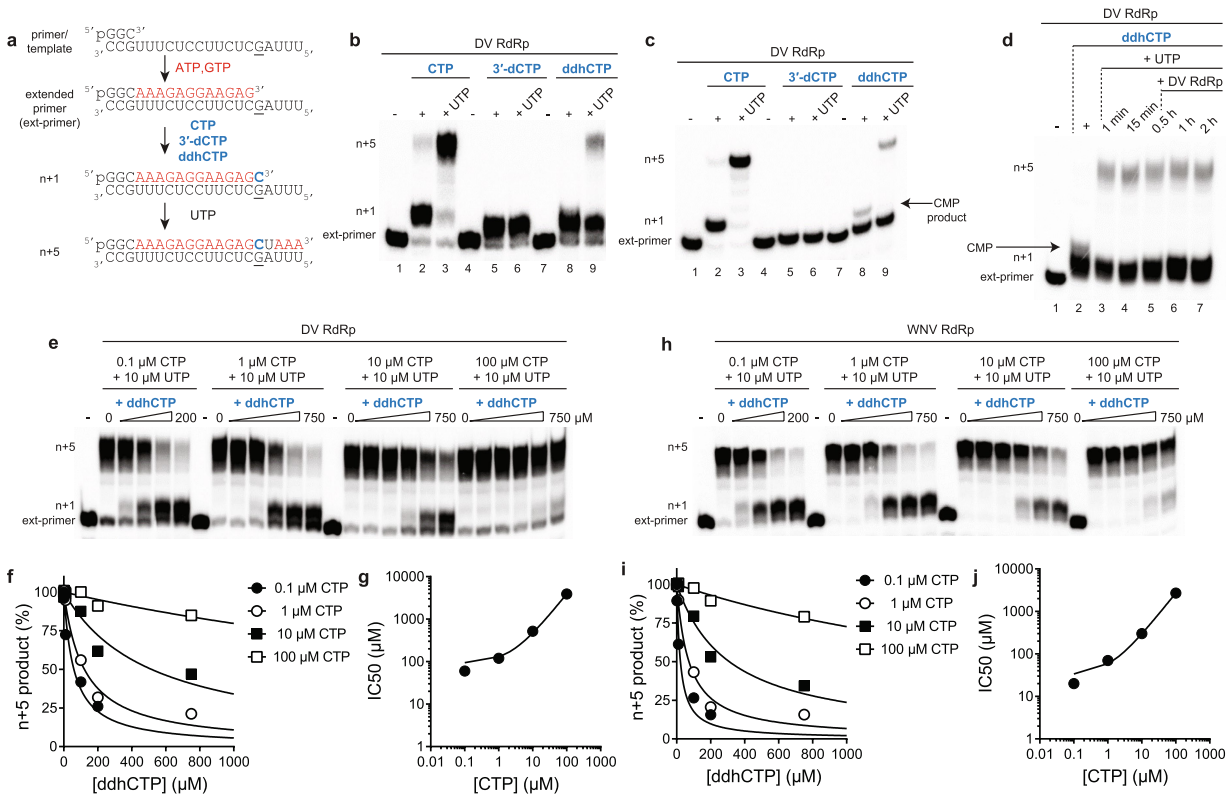
Extended Data Fig. 5 | CMPK2 phosphorylates UDP or CDP and synthetic ddhC can be converted to ddhCTP by cellular machinery. **a**, Formation of trinucleotide species (UTP, CTP or ddhCTP) from mono- and diphosphate species (1 mM UMP, UDP, CMP, CDP and ddhCDP) in the presence of either ATP or GTP as the phosphate donor by 5 μM hCMPK2. **b**, ddhCTP formation in HEK293T cells expressing either Flag–hVIP (N- or C-terminal tags), Flag–hVIP without the N-terminal amphipathic region (delta 1–42), hCMPK2 only, Flag–hVIP (N- or C-terminal tags) and hCMPK2, Flag–hVIP without the N-terminal amphipathic region (delta 1–42) and hCMPK2, a control plasmid or cells only. Only when the tag is on the N terminus of the full-length hVIP is ddhCTP produced at detectable levels. **c**, ddhCTP concentrations from HEK293T suspension cells that were incubated with synthetic ddhC (0 or 1 mM) for 24, 48 or 72 h (see Supplementary Information for details). **d**, ddhCTP concentrations from adherent Vero cells that were incubated with synthetic ddhC (0, 0.3 or 1 mM) for 24, 48 or 72 h (see Supplementary Information for details). All experiments were repeated once. nd, not detectable.



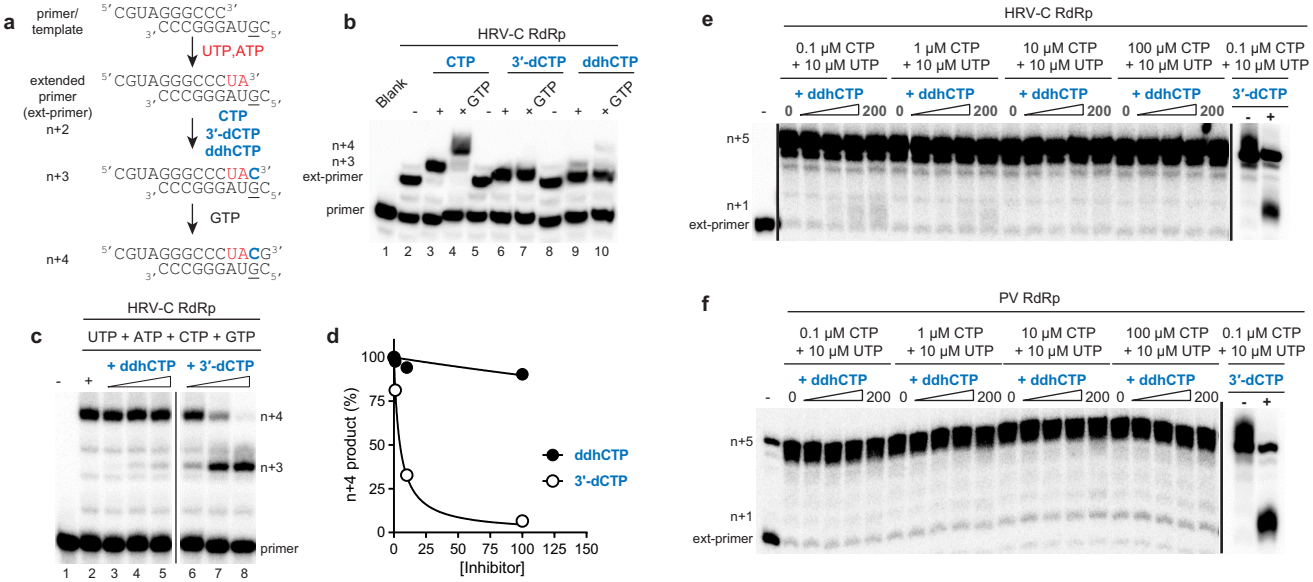
Extended Data Fig. 6 | Cellular concentrations of nucleotides are not affected by viperin expression. a–d, HEK293T cells expressing Flag–hVIP (aqua), Flag–hVIP and hCMPK2 (purple) or cells only (dark blue). Samples were taken at 16, 24, 48 and 72 hours post infection (h.p.i.). Extraction performed with a mixture of acetonitrile, methanol and water (40:40:20 and 0.1 M formic acid). Cellular concentrations were determined using $^{13}C_9{}^{15}N_{15}$–CTP, $^{13}C_{10}{}^{15}N_{10}$–ATP, $^{13}C_{10}N_5$–GTP and $^{13}C_9{}^{15}N_2$–UTP spiked into the extraction mixture at known concentrations and using equations (1) and (2) in Supplementary Information. Analysis of nucleotides ATP (**a**), CTP (**b**), GTP (**c**) and UTP (**d**) did not show statistically significant differences (ns) between Flag–hVIP, Flag–hVIP and hCMPK2 or cells only for any time point. Data are from three biologically independent samples. Statistical significance was determined using a two-way ANOVA (Supplementary Tables 12, 13, 14 and 15). **e**, Ratio of cellular concentrations of ddhCTP to CTP from HEK293T cells expressing Flag–hVIP (aqua) or Flag–hVIP and hCMPK2 (purple); samples were taken at 16, 24, 48 and 72 h post transfection. The overall ratio of ddhCTP to CTP remains constant when only Flag–hVIP is expressed, but the concentration of ddhCTP is boosted significantly relative to CTP when both Flag–hVIP and hCMPK2 are co-expressed (plots are derived from data shown in **c** and Fig. 3a).



Extended Data Fig. 7 | Nucleotide concentrations are not affected during ddhCTP production. a, **b**, Concentrations of ddhCTP (**a**) and CTP, UTP and ATP (**b**) in immortalized macrophage cells (RAW 264.7) grown in serum-free medium in the presence of increasing concentrations of murine IFNα (10 ng ml^{-1}, 50 ng ml^{-1} and 250 ng ml^{-1}). Data are from two biologically independent samples.



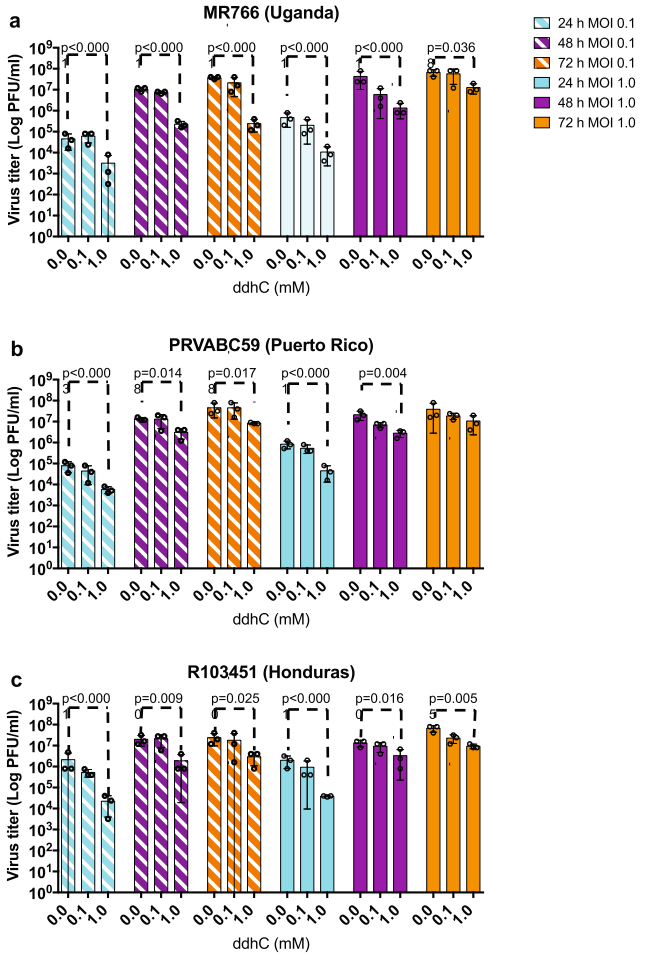
Extended Data Fig. 8 | ddhCTP is used as a substrate by dengue virus and WNV RdRp and chain terminates RNA synthesis. a, Schematic of primer-extension assay for evaluating dengue virus and WNV RdRp activity. **b**, Dengue virus RdRp-catalysed nucleoside incorporation using CTP, 3′-dCTP or ddhCTP as nucleoside triphosphate substrates. Some full-length product was observed in the presence of ddhCTP (more than 99% pure), which is due to residual contaminating CTP that could not be removed. **c**, Reaction products resolved by denaturing PAGE containing 40% formamide showing the trace amount of CTP contaminate in the ddhCTP preparation. These experiments were repeated independently at least four times with similar results. **d**, Longer incubation times and more dengue virus RdRp enzyme does not increase the yield of extended product. **e**, Dengue virus RdRp-catalysed nucleoside incorporation with increasing concentrations of ddhCTP (0, 1, 10, 100, 200 and 750 μM) at varying concentrations of CTP. This experiment was repeated independently three times with similar results. **f**, Plot of the percentage inhibition as a function of ddhCTP concentration at varying concentrations of CTP. Data were fit to a dose–response curve to obtain half-maximum inhibitory concentration (IC_{50}) values of ddhCTP of 60 ± 10, 120 ± 20, 520 ± 90 and 3,900 ± 700 μM at 0.1, 1, 10 and 100 μM CTP, respectively. This experiment was repeated at least three times with similar results. The total sample size is 24. The error reported is the standard error from the fit of the data to a dose–response curve. **g**, Plot of IC_{50} values as a function of CTP concentration. The data were fit to a line with a slope of 38 ± 1 and an intercept of 91 ± 25. The error reported is the standard error from the fit of the data to a line. **h**, WNV RdRp-catalysed nucleoside incorporation with increasing concentrations of ddhCTP (0, 1, 10, 100, 200 and 750 μM) at varying concentrations of CTP. This experiment was repeated at least three times with similar results. **i**, Plot of the percentage inhibition as a function of ddhCTP concentration at varying concentrations of CTP. Data were fit to a dose–response curve to obtain IC_{50} values of ddhCTP of 20 ± 10, 70 ± 10, 300 ± 40 and 2,700 ± 300 μM at 0.1, 1, 10 and 100 μM CTP, respectively. The total sample size is 24. The error reported is the standard error from the fit of the data to a dose–response curve. **j**, Plot of IC_{50} values as a function of CTP concentration. The data were fit to a line with a slope of 27 ± 1 and an intercept of 31 ± 8. Both of these results demonstrate that once ddhCMP is incorporated, it effectively terminates synthesis and that the small amount of extended product is from a trace amount of CTP contamination. The error reported is the standard error from the fit of the data to a line.



g

RdRp	IC50[a]	Probability[b]	% full-length genome[c]
DV RdRp	3×10^{4}	0.011	3×10^{-11}
WNV RdRp	2×10^{4}	0.015	4×10^{-17}
HRV-C RdRp	3×10^{6}	1×10^{-4}	86
PV RdRp	$\geq 3 \times 10^{7}$	$\geq 1 \times 10^{-5}$	99

Extended Data Fig. 9 | HRV-C and poliovirus RdRp are poorly inhibited by ddhCTP. a, Schematic of primer extension assay for evaluating HRV-C RdRp activity. **b**, HRV-C RdRp-catalysed nucleoside incorporation using CTP, 3′-dCTP or ddhCTP as nucleoside triphosphate substrates. These experiments were repeated independently at least four times with similar results. **c**, Increasing concentrations of ddhCTP does not efficiently inhibit HRV-C RdRp-catalysed RNA synthesis. HRV-C RdRp-catalysed nucleoside incorporation in the presence of increasing concentrations of either ddhCTP or 3′-dCTP. These experiments were repeated independently at least five times with similar results. **d**, Plot of the percentage inhibition as a function of either ddhCTP or 3′-dCTP concentration. Data were fit to a dose–response curve to obtain IC_{50} values of 900 ± 300 μM for ddhCTP and 5 ± 1 μM for 3′-dCTP. The total sample size is eight. The error reported is the standard error from the fit of the data to a dose–response curve. **e**, **f**, HRV-C (**e**) and poliovirus (**f**) RdRp-catalysed nucleoside incorporation with increasing concentrations of ddhCTP (0, 1, 10, 100 and 200 μM) at varying concentrations of CTP. Reactions were performed with the trinucleotide primer, 5′-pGGC, and 20-nt RNA template as described for dengue virus and WNV RdRp to directly compare results with HRV-C and poliovirus RdRp. At the highest concentration of ddhCTP, only approximately 2% inhibition was observed for HRV-C RdRp at 0.1 and 1 μM CTP. The IC_{50} values at 0.1 and 1 μM CTP are estimated to be approximately 10,000 and 20,000 μM ddhCTP, respectively. These values are three orders of magnitude higher than those obtained for dengue virus and WNV RdRp. Reactions in the presence of 3′-dCTP (200 μM) are shown as a control for inhibition. These experiments were repeated independently at least four times with similar results. **g**, Efficiency of incorporation and inhibition of viral RdRps. Footnotes: [a]Calculated for ddhCTP in direct competition with CTP (800 μM) using the linear equations obtained from the fit of the data shown in **g** and **j**. For HRV-C, the IC_{50} value was estimated to be two orders magnitude greater than that calculated for dengue virus and WNV RdRps as evidenced from the data shown in **d**. [b]Calculated for a ddhCTP concentration of 350 μM using the following equation: probability = [ddhCTP]/([ddhCTP] + IC_{50}). [c]Calculated using the following equation: $\text{full} - \text{lengthgenome}(\%) = 100 \times (1 - \text{probability})^{C_n}$; in which C_n is the number of cytidine residues in the viral genome with values of 2,200, 2,497, 1,565 and 1,737 for dengue virus, WNV, HRV-C and poliovirus respectively.



Extended Data Fig. 10 | ddhC reduces virus release of three different ZIKV isolates. Vero cells were treated with different concentrations of ddhC (0, 0.1 and 1 mM) for 24 h. After this 24-h period, the medium over cells was removed and cells were infected with fresh medium that contained the original concentrations of ddhC (0, 0.1 and 1 mM) and one of three strains of ZIKV; African strain MR766 (Uganda 1947), PRVABC59 (Puerto Rico; 2015) or R103451 (Honduras; 2016, GenBank: KX262887). After three hours of ZIKV infection, virus inoculum was removed and cells were treated with fresh medium that contained the original concentrations of ddhC (0, 0.1 and 1 mM). Virus samples were collected and the medium over cells was replaced with fresh medium that contained the original concentrations of ddhC at 24, 48 and 72 h.p.i. Viral titres at 24, 48 and 72 h.p.i. were determined using the plaque assay. **a**–**c**, Effect of ddhC on three different ZIKV isolates: MR766 (Uganda 1947) (**a**), PRVABC59 (Puerto Rico; 2015) (**b**) or R103451 (Honduras; 2016) (**c**). Analysis of ZIKV titres indicates that 1 mM ddhC inhibits all three ZIKV isolates compared to 0 mM ddhC. However, reduction in virus titre is more prominent at 24 h.p.i. and 48 h.p.i. compared to 72 h.p.i. when using an MOI of 1.0. The antiviral effect of ddhC is more prominent at an MOI of 0.1. Data are mean ± s.d. from three biologically independent samples, *P* values from a two-way ANOVA with Dunnett's post hoc analysis.

LETTER

https://doi.org/10.1038/s41586-018-0241-9

Cryo–EM structure of the serotonin 5–HT_{1B} receptor coupled to heterotrimeric G_o

Javier García–Nafría[1,2], Rony Nehmé[1,2], Patricia C. Edwards[1] & Christopher G. Tate[1]*

G-protein-coupled receptors (GPCRs) form the largest family of receptors encoded by the human genome (around 800 genes). They transduce signals by coupling to a small number of heterotrimeric G proteins (16 genes encoding different α-subunits). Each human cell contains several GPCRs and G proteins. The structural determinants of coupling of G_s to four different GPCRs have been elucidated[1–4], but the molecular details of how the other G-protein classes couple to GPCRs are unknown. Here we present the cryo-electron microscopy structure of the serotonin 5-HT_{1B} receptor (5-HT_{1B}R) bound to the agonist donitriptan and coupled to an engineered G_o heterotrimer. In this complex, 5-HT_{1B}R is in an active state; the intracellular domain of the receptor is in a similar conformation to that observed for the β_2-adrenoceptor (β_2AR)[3] or the adenosine A_{2A} receptor (A_{2A}R)[1] in complex with G_s. In contrast to the complexes with G_s, the gap between the receptor and the Gβ-subunit in the G_o–5-HT_{1B}R complex precludes molecular contacts, and the interface between the Gα-subunit of G_o and the receptor is considerably smaller. These differences are likely to be caused by the differences in the interactions with the C terminus of the G_o α-subunit. The molecular variations between the interfaces of G_o and G_s in complex with GPCRs may contribute substantially to both the specificity of coupling and the kinetics of signalling.

Heterotrimeric G proteins can be divided into four subfamilies[5], G_s, $G_{i/o}$, G_q and $G_{12/13}$, each containing α, β and γ-subunits. When an agonist binds to a GPCR, the receptor couples to a G protein, predominantly through the α-subunit (Gα); there are relatively few contacts with the β-subunit and none with the γ-subunit. The overall structure of α-subunits in the inactive GDP-bound state is highly conserved[6], and they undergo similar conformational changes upon coupling to GPCRs[7]. This is characterized by a disorder-to-order transition of the C-terminal half of the α5-helix, which adopts an α-helical conformation upon binding in the cytoplasmic cleft of an activated GPCR[8]. The key role of the α5-helix in G-protein coupling to GPCRs has been confirmed through mutagenesis studies, which are consistent with the α5-helix accounting for 70% of the interactions between Gα and β_2AR[3]. The amino acid sequences of the C terminus of the α5-helix are highly conserved within G-protein subfamilies, but are distinct between different subfamilies. The key role of this region in determining specificity is indicated by the ability to change the coupling specificity of a G protein by mutating the C-terminal region to match that of a different G protein[9]. However, other regions of the α-subunit also contribute to specificity, and mutations often do not map directly to specificity when transferred from one G protein to another[10]. Many receptors couple to more than one G protein, and this coupling can appear different depending upon whether coupling is measured dynamically or in an end-point assay[11]. The reported structure of β_2AR coupled to heterotrimeric G_s changed the model of how G proteins couple to and are activated by GPCRs[3], but it does not address the issue of G-protein specificity. We have therefore determined the structure of 5-HT_{1B}R in complex with heterotrimeric G_o to enable comparisons of receptor coupling to G_s and G_o.

There are 13 GPCRs in the serotonin receptor family[12], all of which are expressed in the central nervous system where they have key roles in all aspects of behaviour[13]. Structures of several 5-HT receptors have been determined in either inactive or active-intermediate states[14–16]. 5-HT_{1B}R binds the agonist donitriptan with high affinity and couples to $G_{i/o}$[17]. G_o is the most abundant G protein in the brain and an engineered G_o, mini-G_o, was developed to form a heterotrimer with the Gβ and Gγ subunits, which can bind and stabilize the agonist-activated 5-HT_{1B}R[10]. We expressed and purified these proteins and assembled them into a complex containing 5-HT_{1B}R, donitriptan, mini-G_o, β_1 and γ_2 subunits (see Methods). The purified complex was vitrified on electron microscopy grids and the structure was determined by cryo-electron microscopy (cryo-EM) and single-particle analysis to an overall resolution of 3.8 Å (Extended Data Figs. 1–4, Extended Data Table 1), with clear density for the majority of side chains and the agonist donitriptan (Fig. 1, Extended Data Fig. 2). Donitriptan occupies the orthosteric binding site, and the serotonin-like moiety of the ligand binds in a region analogous to that identified for the native agonists adrenaline[18] and adenosine[19] (Fig. 1, Extended Data Fig. 5). Donitriptan binds 5-HT_{1B}R in a different mode to the ergot family of alkaloids, such as ergotamine and dihydroergotamine[15] (Fig. 1). The donitriptan-binding site is formed by amino acid residues in transmembrane helices 3, 5, 6 and 7 (H3, H5, H6 and H7) and extends into the extracellular region to make contacts with H6, H7 and extracellular loop 2 (ECL2). Donitriptan is bound primarily by van der Waals contacts and limited polar interactions with Thr134[3.37] and Asp129[3.32].

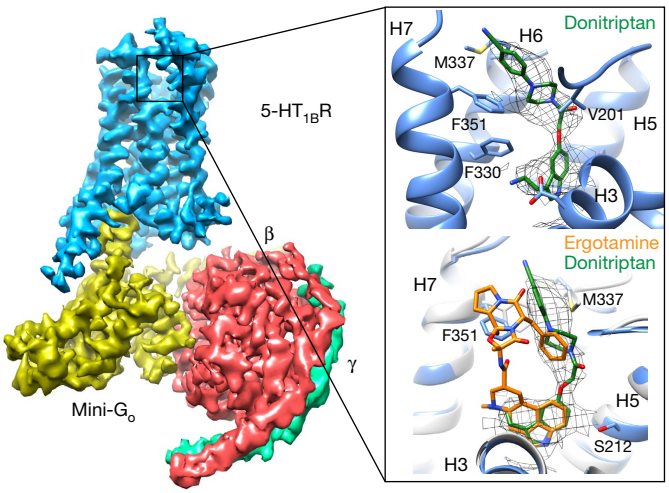


Fig. 1 | Overall cryo-EM reconstruction of the 5-HT_{1B}R–G_o heterotrimer complex. The density of the cryo-EM map (sharpened with a *B* factor of −200) is coloured according to the subunit. The inset shows the orthosteric binding pocket in 5-HT_{1B}R (light blue) with donitriptan depicted as sticks (green, carbon) and its density in the cryo-EM map. The lower panel shows a superposition of ergotamine-bound 5-HT_{1B}R (pale grey, PDB code 4IAR)[15] and donitriptan-bound 5-HT_{1B}R (pale blue). Donitriptan (green, carbon) and ergotamine (orange, carbon) are depicted as sticks.

[1]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge, CB2 0QH, UK. [2]These authors contributed equally: Javier García-Nafría, Rony Nehmé. *e-mail: cgt@mrc-lmb.cam.ac.uk

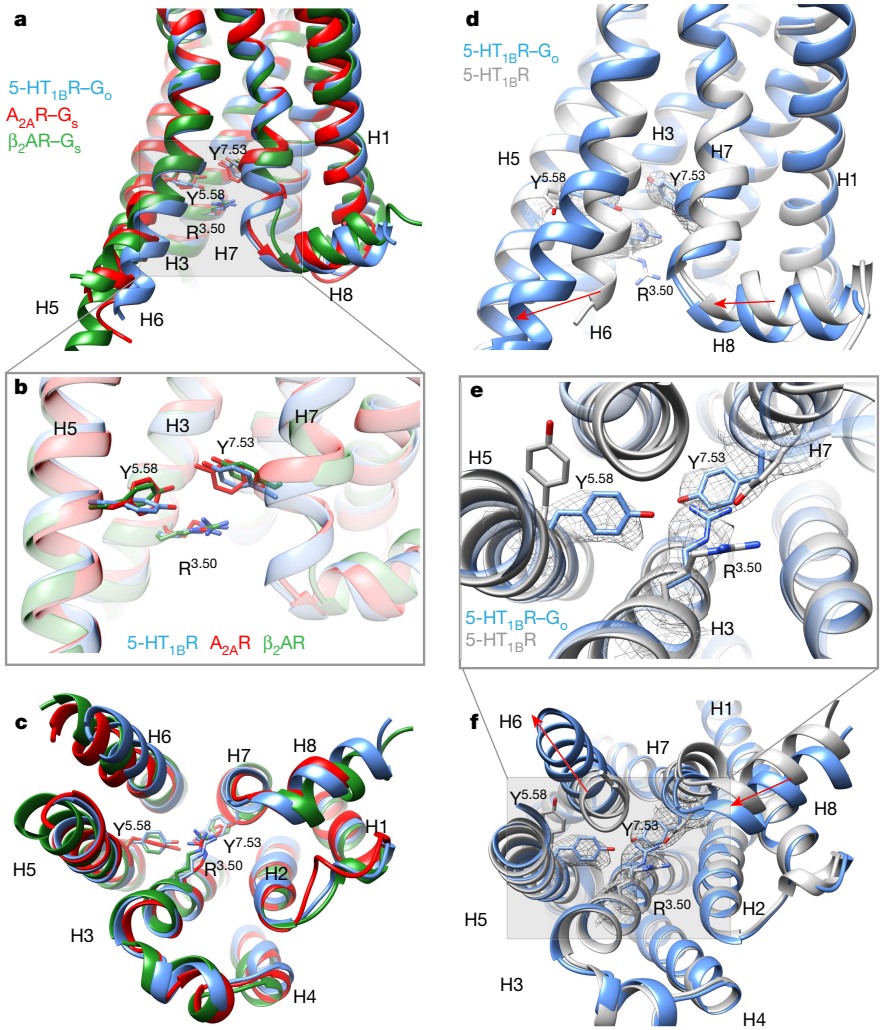


Fig. 2 | G_o-coupled 5-HT_{1B}R is in an active conformation. **a**–**c**, Superposition of G-protein-bound receptors: based on H3, H5 and H6: 5-HT_{1B}R (blue), A_{2A}R (red)[1] and β_2AR (green)[3]. Key amino acid residues involved in receptor activation are displayed as sticks. **d**–**f**, Superposition of G_o-coupled 5-HT_{1B}R (blue) and the active-intermediate state of 5-HT_{1B}R bound to ergotamine (pale grey), based on alignment of the whole receptor. Conformational changes involved in receptor activation are highlighted (red arrows) and key residues are shown as sticks with the density from the cryo-EM map (mesh). **a** and **d**, view parallel to the plane of the membrane; **b** and **e**, enlarged view of the conserved core of the receptors; **c** and **f**, view from the cytoplasmic face of the membrane.

(unless otherwise noted, superscripts indicate Ballesteros–Weinstein numbering for GPCRs[20]), similar to the binding mode of ergotamine. However donitriptan and ergotamine bind on opposite faces of the binding pocket, and different rotamers of Phe351$^{7.35}$ and Met337$^{6.58}$ accommodate the two ligands. The resolution at the ligand binding pocket is slightly lower than for the core of the complex and there are limitations in the interpretation of the experimental data (see Methods and Extended Data Fig. 2).

The overall conformation of 5-HT_{1B}R in the cryo-EM structure is consistent with the receptor being in a fully active state. Superposition with the active states of β_2AR[3] and A_{2A}R[1] shows a high degree of conservation of the cytoplasmic region of the receptors (Fig. 2). In addition, rotamers of key conserved amino acid residues (Pro$^{5.50}$, Ile$^{3.40}$, Phe$^{6.44}$)[8] are almost identical to those in the active-state structures, suggesting that 5-HT_{1B}R is also in a fully active state. The structure of a 5-HT_{1B}R–BRIL fusion bound to the agonist ergotamine has been determined and was suggested to be in an active intermediate state owing to the partial movement of H6 and partial rotamer changes of key amino acid residues[15]. Comparison of the G_o-coupled 5-HT_{1B}R with the ergotamine-bound 5-HT_{1B}R shows an 8 Å shift of the cytoplasmic end of H6 (C_α of Lys311), an inward shift of H7 and H8 by 2 Å (C_α of Glu374) and changes in rotamer of Arg$^{3.50}$, Tyr$^{5.58}$ and Tyr$^{7.53}$ (Fig. 2). Of note, the conformation of the extracellular region of the receptor that forms the orthosteric binding pocket does not exhibit any marked change in conformation in the transition from the active intermediate state to the active G-protein-coupled state. This resembles the transition of the A_{2A}R[21], but different to that observed in β_2AR[3], which has a different energy landscape[22].

The overall architecture of 5-HT_{1B}R coupled to G_o is similar to that of the β_2AR–G_s and A_{2A}R–G_s complexes[1,3], but there are critical differences in the details. The interface between 5-HT_{1B}R and G_o consists of 9 amino acid residues from 5-HT_{1B}R and 13 from the α-subunit of G_o. This compares with 24 residues from β_2AR and 20 in A_{2A}R that make contact with 17 and 22 residues in $G\alpha_s$, respectively (Fig. 3). The surface area of G_o in contact with 5-HT_{1B}R is 822 Å^2, whereas the areas of G_s in contact with β_2AR and A_{2A}R are 1260 Å^2 and 1135 Å^2, respectively. All the contacts made by G_o with the receptor are through residues in the α5-helix. By contrast, contacts made by $G\alpha_s$ to β_2AR and A_{2A}R also involve regions in S1, S2–S3 and H4–S6 (Extended Data Fig. 6). The overall conformations of $G\alpha_s$ and $G\alpha_o$ are very similar when coupled to the receptors (Extended Data Fig. 7). However, alignment of the cytoplasmic regions of 5-HT_{1B}R, β_2AR and A_{2A}R shows that the α5-helix of $G\alpha_o$ is positioned differently within the receptor compared to $G\alpha_s$ (Fig. 3, Extended Data Fig. 7). There is a 9° or 11° tilt of the N-terminal end of the α5-helix away from the plane of the membrane when compared to G_s coupled to β_2AR or A_{2A}R, respectively

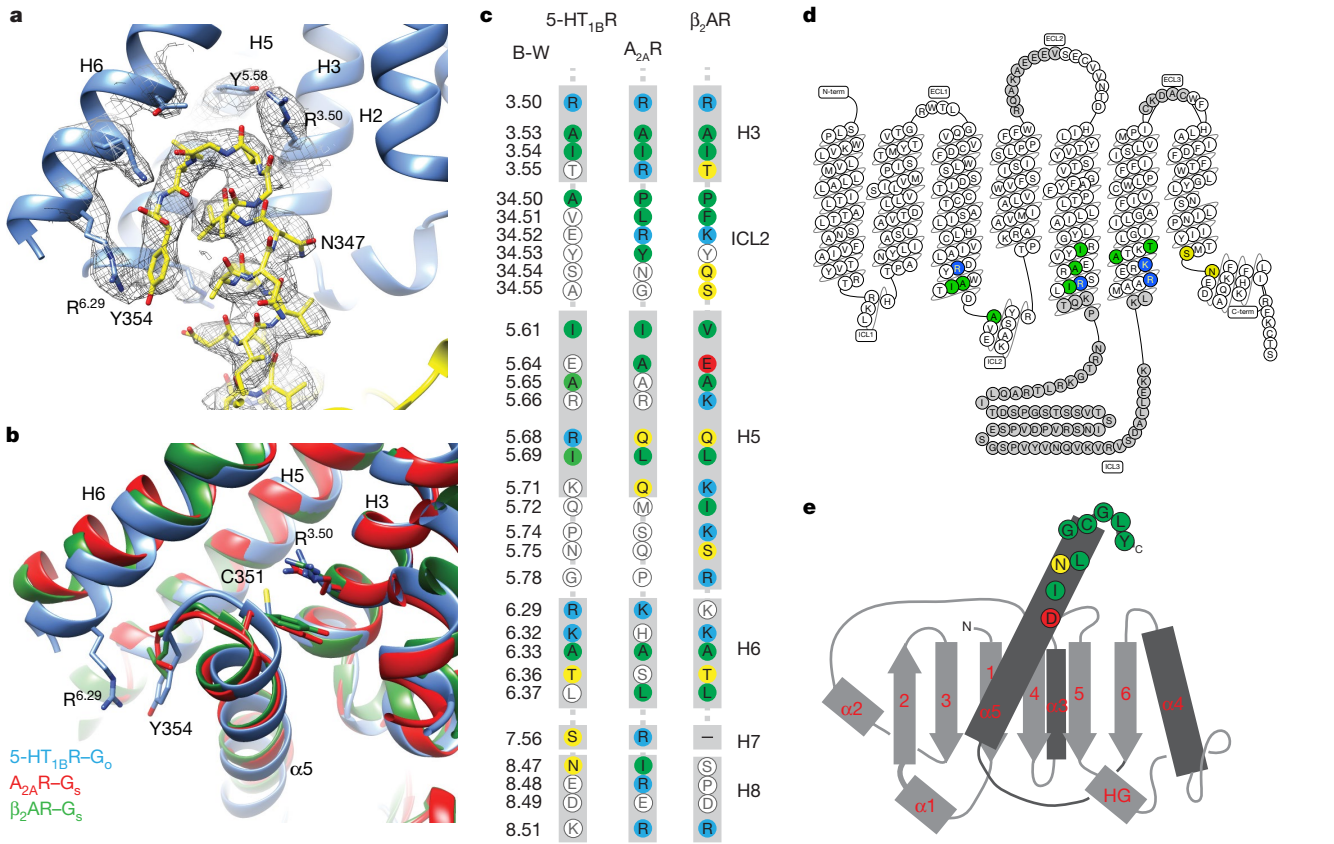


Fig. 3 | G_o coupling to the 5-HT_{1B}R. a, C-terminal end of $G\alpha_o$ (yellow sticks) inserted into the cytoplasmic cleft of 5-HT_{1B}R (blue cartoon). Cryo-EM density is depicted as a mesh. **b**, Superposition of 5-HT_{1B}R (blue), A_{2A}R (red)[1] and β_2AR (green)[3] based on alignment of H3, H5 and H6. The different poses of the C termini of G_s and G_o coupled to the respective receptors are shown. **c**, Amino acid residues in 5-HT_{1B}R, A_{2A}R and β_2AR that make contact with the respective $G\alpha$ subunits that they are coupled to are shown in colours that reflect biophysical properties of the residues. Green, hydrophobic; yellow, hydrophilic, red, acidic; blue, basic. Residues coloured in white do not make contact to the relevant $G\alpha$. The amino acid alignment was created in GPCRdb[30] and secondary structural elements and the Ballesteros–Weinstein numbers[20] are shown. **d**, Snake plot of 5-HT_{1B}R created in GPCRdb with residues making contact with $G\alpha_o$ coloured according to their biophysical properties, coded as in **c**. Regions in grey were disordered in the cryo-EM map. **e**, Cartoon of secondary structural elements in $G\alpha_o$ and amino acid residues that make contact with 5-HT_{1B}R are depicted and coloured according to their biophysical properties as in **c**.

(the pivot point is in the region of I344–N347 in $G\alpha_o$). The different tilt angles are probably the result of the different positions of the approximately eight C-terminal amino acid residues of the α5-helices that are located within the receptor. This region contains the major determinants of specificity for different G proteins[7]. The final four amino acid residues in the α5-helix of $G\alpha_s$ are $Y^{H5.23}ELL^{H5.26}$; the equivalent residues in $G\alpha_o$ are $C^{H5.23}GLY^{H5.26}$ (superscripts refer to the CGN numbering system[7]). These residues form a 'wavy hook' structure at the end of the α5-helix in $G\alpha_s$. In $G\alpha_s$, the π-electrons of $Tyr^{H5.23}$ form extensive contacts with the positively charged $Arg^{3.50}$, which forms the boundary between the cytoplasmic cleft where the α5-helix binds and the hydrophobic core of the receptor[21]. Similarly, in G_o, $Cys351^{H5.23}$ interacts with $Arg147^{3.50}$, although only through van der Waals interactions. Therefore, the α5-helix of both G_o and G_s penetrate GPCRs to the same degree. In contrast to $G\alpha_s$, the single amino acid residue in $G\alpha_o$ that makes most contacts with the receptor is the C-terminal $Tyr354^{H5.26}$, the side chain of which stacks against $Arg308^{6.29}$ in 5-HT_{1B}R and also makes a weak polar interaction with the same residue. In $G\alpha_s$, the terminal amino acid $Leu394^{H5.26}$ makes only very few contacts with β_2AR and is disordered in the A_{2A}R–G_s structure. In the A_{2A}R–mini-G_s crystal structure, there are extensive contacts between $Glu394^{H5.24}$ and three Arg residues in the H7/H8 region of the receptor[21]; the equivalent residue in G_o is $Gly352^{H5.24}$, which makes only minor contacts with the receptor. Although it appears from the cryo-EM structure that all the major contacts between 5-HT_{1B}R and $G\alpha_o$ are mediated by the α5-helix of $G\alpha_o$, there is weak density for H5 and H6 of 5-HT_{1B}R that extends towards the α4 helix in $G\alpha_o$ (Extended Data Fig. 1). It is known that mutations in the α4 helix can affect coupling to 5-HT_{1B}R[23], but it is unclear from the structure whether this is because direct contacts to the receptor are absent from the mutated α4 helix, or because there is a secondary effect of the mutation on the structure of $G\alpha_o$.

The determinants of coupling specificity of G proteins are found predominantly at the C terminus of $G\alpha$ in the α5-helix and the wavy hook. The architecture of this region is virtually identical in $G\alpha_s$ and $G\alpha_o$, but the differences in amino acid sequence (Extended Data Figs 6, 8) result in $G\alpha_s$ being bulkier than $G\alpha_o$ in the terminal five residues (Extended Data Fig. 9). This may be sufficient to prevent coupling of G_s to some G_o-specific GPCRs as the narrower crevice in these GPCRs may exclude the bulkier C terminus of G_s. Conversely, the wider crevice in G_s-coupled receptors may allow coupling of G_o, provided that there are suitable residues lining the crevice to form a good interface. This last caveat raises the problem of predicting G-protein-coupling specificity. Although the structure and mechanism of GPCRs are highly conserved[8,24], human GPCRs show considerable sequence heterogeneity; therefore, there is little or no specific amino acid conservation correlating with the subtype of G protein that a receptor couples to[25]. In addition, there is potential for different GPCR conformations[26], which suggests that the mode of G-protein coupling could be different between different receptors. This is the case in the complex of transducin peptide and opsin[27], in which the α5-helix is tilted by around 30° in comparison to the α5-helix of G_o, even though the G proteins are in the same family. More structures need to be determined to evaluate the diversity of G-protein coupling.

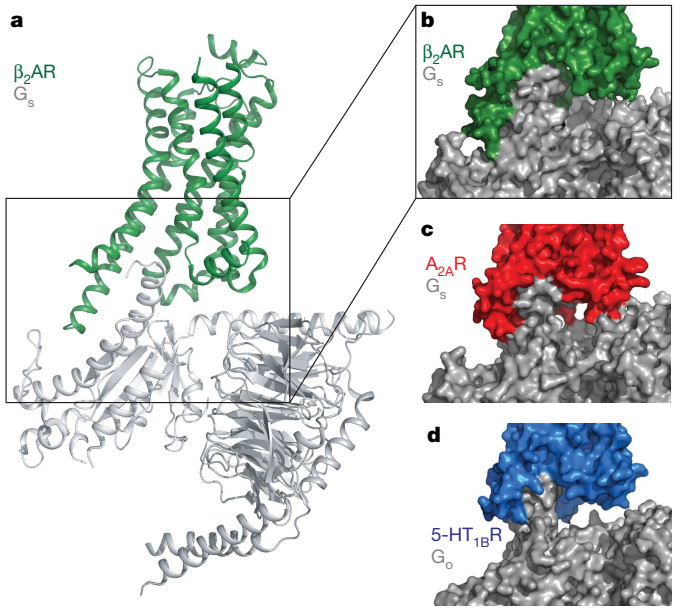


Fig. 4 | Comparison of G_s and G_o coupling. a, Cartoon of β_2AR (green) coupled to G_s (PDB code 3SN6)[3]. The α-helical domain of Gα has been removed for clarity. **b**–**d**, Surface-rendered views of the interface between receptor and G protein: β_2AR (green) and G_s (**b**); $A_{2A}R$[1] (red) and G_s (**c**); 5-HT_{1B}R (blue) and G_o (**d**).

The specific differences in packing at the C terminus of G_o compared to G_s have a disproportionate effect on the whole G protein owing to their amplification as a result of the different insertion angle of the α5-helix. This results in a change in the tilt of the whole G protein, which moves away from the plane of the membrane and results in a gap between the rest of the G protein and 5-HT_{1B}R. Therefore, there are no contacts between 5-HT_{1B}R and Gβ subunits, and the only contacts made to Gα are with the α5-helix. This is in marked contrast to the relatively close packing of G_s to both $A_{2A}R$ and β_2AR (Fig. 4). Given that the mechanism of GPCR[24] and G-protein activation is conserved[7], it is likely that the small interface between 5-HT_{1B}R and G_o is a common feature of receptor coupling with $G_{i/o}$ family, and will be seen in other GPCRs that are activated by diffusible ligands. A likely consequence of the small interface in the receptor–$G_{i/o}$ complex is that $G_{i/o}$ may have a faster rate of dissociation than G_s in the same GPCR. The kinetics of the steps in GPCR signalling pathways are thought to have a profound effect on which particular signalling event results from agonist binding to a receptor in a specific cell type[28,29]. A combination of structural data and kinetic analyses will be essential to unravel the complexities of this system.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0241-9

Received: 16 February 2018; Accepted: 22 May 2018;
Published online: 20 June 2018

1. García-Nafría, J., Lee, Y., Bai, X., Carpenter, B. & Tate, C. G. Cryo-EM structure of the adenosine A_{2A} receptor coupled to an engineered heterotrimeric G protein. *eLife* **7**, e35946 (2018).
2. Liang, Y. L. et al. Phase-plate cryo-EM structure of a class B GPCR–G protein complex. *Nature* **546**, 118–123 (2017).
3. Rasmussen, S. G. et al. Crystal structure of the β_2 adrenergic receptor–G_s protein complex. *Nature* **477**, 549–555 (2011).
4. Zhang, Y. et al. Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature* **546**, 248–253 (2017).
5. Syrovatkina, V., Alegre, K. O., Dey, R. & Huang, X. Y. Regulation, signaling, and physiological functions of G proteins. *J. Mol. Biol.* **428**, 3850–3868 (2016).
6. Oldham, W. M. & Hamm, H. E. Structural basis of function in heterotrimeric G proteins. *Q. Rev. Biophys.* **39**, 117–166 (2006).
7. Flock, T. et al. Universal allosteric mechanism for Gα activation by GPCRs. *Nature* **524**, 173–179 (2015).
8. Venkatakrishnan, A. J. et al. Molecular signatures of G protein-coupled receptors. *Nature* **494**, 185–194 (2013).
9. Oldham, W. M. & Hamm, H. E. Heterotrimeric G protein activation by G protein-coupled receptors. *Nat. Rev. Mol. Cell Biol.* **9**, 60–71 (2008).
10. Nehme, R. et al. Mini-G proteins: novel tools for studying GPCRs in their active conformation. *PLoS One* **12**, e0175642 (2017).
11. Masuho, I. et al. Distinct profiles of functional discrimination among G proteins determine the actions of G protein-coupled receptors. *Sci. Signal.* **8**, ra123 (2015).
12. McCorvy, J. D. & Roth, B. L. Structure and function of serotonin G protein-coupled receptors. *Pharmacol. Ther.* **150**, 129–142 (2015).
13. Berger, M., Gray, J. A. & Roth, B. L. The expanded biology of serotonin. *Annu. Rev. Med.* **60**, 355–366 (2009).
14. Wacker, D. et al. Structural features for functional selectivity at serotonin receptors. *Science* **340**, 615–619 (2013).
15. Wang, C. et al. Structural basis for molecular recognition at serotonin receptors. *Science* **340**, 610–614 (2013).
16. Yin, W. et al. Crystal Structure of the human 5-HT_{1B} serotonin receptor bound to an inverse agonist. *Cell Discovery* **4**, 12 (2018).
17. Albert, P. R. & Tiberi, M. Receptor signaling and structure: insights from serotonin-1 receptors. *Trends Endocrinol. Metab.* **12**, 453–460 (2001).
18. Ring, A. M. et al. Adrenaline-activated structure of β_2-adrenoceptor stabilized by an engineered nanobody. *Nature* **502**, 575–579 (2013).
19. Lebon, G. et al. Agonist-bound adenosine A_{2A} receptor structures reveal common features of GPCR activation. *Nature* **474**, 521–525 (2011).
20. Ballesteros, J.A., Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors *Methods Neurosci.* **25**, 366–428 (1995).
21. Carpenter, B., Nehme, R., Warne, T., Leslie, A. G. & Tate, C. G. Structure of the adenosine A_{2A} receptor bound to an engineered G protein. *Nature* **536**, 104–107 (2016).
22. Lebon, G., Warne, T. & Tate, C. G. Agonist-bound structures of G protein-coupled receptors. *Curr. Opin. Struct. Biol.* **22**, 482–490 (2012).
23. Bae, H., Cabrera-Vera, T. M., Depree, K. M., Graber, S. G. & Hamm, H. E. Two amino acids within the α4 helix of $G\alpha_{i1}$ mediate coupling with 5-hydroxytryptamine$_{1B}$ receptors. *J. Biol. Chem.* **274**, 14963–14971 (1999).
24. Venkatakrishnan, A. J. et al. Diverse activation pathways in class A GPCRs converge near the G protein-coupling region. *Nature* **536**, 484–487 (2016).
25. Flock, T. et al. Selectivity determinants of GPCR–G protein binding. *Nature* **545**, 317–322 (2017).
26. Kobilka, B. K. & Deupi, X. Conformational complexity of G protein-coupled receptors. *Trends Pharmacol. Sci.* **28**, 397–406 (2007).
27. Scheerer, P. et al. Crystal structure of opsin in its G protein-interacting conformation. *Nature* **455**, 497–502 (2008).
28. Grundmann, M. & Kostenis, E. Temporal bias: time-encoded dynamic GPCR signaling. *Trends Pharmacol. Sci.* **38**, 1110–1124 (2017).
29. Lane, J. R., May, L. T., Parton, R. G., Sexton, P. M. & Christopoulos, A. A kinetic view of GPCR allostery and biased agonism. *Nat. Chem. Biol.* **13**, 929–937 (2017).
30. Isberg, V. et al. GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **44**, D356–D364 (2016).

Acknowledgements This work was funded by a grant from the European Research Council (EMPSI 339995), Heptares Therapeutics and core funding from the Medical Research Council (MRC U105197215). We thank J. Espinosa and L. Renault for their help with data collection at NeCEN; S. Scheres and P. da Fonseca for useful discussions and C. Savva and G. Cannone for microscopy technical support.

Reviewer information: *Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.N. performed receptor expression, purification and complex formation. P.C.E. expressed and purified mini-G_o and $G\beta\gamma$. R.N. and J.G.-N. performed cryo-grid preparation. J.G.-N. performed cryo-EM data collection, data processing and model building. J.G.-N. and C.G.T. carried out structure analysis and manuscript preparation. C.G.T. analysed data and managed the project. The manuscript was written by C.G.T and J.G.-N., and included contributions from all the authors.

Competing interests C.G.T. is a shareholder, consultant and member of the Scientific Advisory Board of Heptares Therapeutics, who also partly funded this work.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0241-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0241-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.G.T.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Expression and purification of 5-HT$_{1B}$R. N-terminally truncated wild-type human 5-HT$_{1B}$R (residues 34–390) was modified to contain a C-terminal histidine tag (His$_{10}$) and TEV protease cleavage site[10]. The L138W$^{3.41}$ mutation was introduced to increase thermostability. Recombinant baculoviruses expressing 5-HT$_{1B}$R were prepared using the flashBAC ULTRA system (Oxford Expression Technologies). *Trichoplusia ni* cells (Expression Systems) were grown in suspension in ESF921 media (Expression Systems) to a density of 3×10^6 cells/ml, infected with 5-HT$_{1B}$R baculovirus and incubated for 48 h. Cells were harvested and membranes prepared by two ultracentrifugation steps in 20 mM HEPES pH7.5, 1 mM EDTA, 1 mM PMSF. Membranes were resuspended finally in 20 mM HEPES pH7.5, 500 mM NaCl, 5 mM $MgCl_2$, 10 mM imidazole and Complete protease inhibitors (Roche) and flash frozen in liquid nitrogen and stored at −80 °C.

Membranes from 2 l of cells were solubilised with 2% *n*-decyl-β-D-maltopyranoside (DM) on ice for 30 min in the presence of 1 μM donitriptan hydrochloride. The sample was clarified by ultracentrifugation and loaded onto a 5 ml Ni-NTA column (Generon). The column was equilibrated and sample was loaded in buffer A (20 mM HEPES pH 7.5, 500 mM NaCl, 1 mM $MgCl_2$, 50 mM imidazole, 1 μM donitriptan hydrochloride, 0.15% DM), and eluted with buffer B (20 mM HEPES pH 7.5, 100 mM NaCl, 1 mM $MgCl_2$, 300 mM imidazole, 1 μM donitriptan hydrochloride, 0.15% DM). The eluate was concentrated using a 50-kDa cut-off Amicon centrifugal ultrafiltration unit (Millipore), and exchanged into desalting buffer (20 mM HEPES pH 7.5, 100 mM NaCl, 1 mM $MgCl_2$, 1 μM donitriptan hydrochloride, 0.15% DM) using a PD10 column (GE Healthcare). Then, 2.5 mg TEV protease was added, and the sample was incubated on ice overnight. TEV protease was removed by negative purification on Ni^{2+}-NTA resin. The sample was concentrated to ~1 ml and loaded onto a Superdex 200 column (GE Healthcare). Peak fractions corresponding to monomers of receptors were pooled and concentrated. A typical yield was 1–2 mg pure 5-HT$_{1B}$R per litre of culture.

Formation of a 5-HT$_{1B}$R–heterotrimeric mini-G$_o$ complex. Purified 5-HT$_{1B}$R was mixed with a 1.2-fold molar excess of mini-G$_{o1}\beta_1\gamma_2$ in the presence of apyrase (0.2 U/ml) and the mixture was incubated on ice overnight[10]. The sample was loaded on to a Superdex 200 column. Peak fractions containing the 5-HT$_{1B}$R–mini-G$_{o1}\beta_1\gamma_2$ complex were pooled and concentrated to 4 mg/ml.

Cryo-grid preparation and data collection. Cryo-EM grids were prepared by applying 3 μl sample (at a protein concentration of 2.2 mg/ml) on glow-discharged holey gold grids (Quantifoil Au 1.2/1.3 300 mesh). Excess sample was removed by blotting with filter paper for 3–4 s before plunge-freezing in liquid ethane using a FEI Vitrobot Mark IV at 100% humidity and 4 °C. Images were collected on a FEI Titan Krios microscope at 300 kV using a Falcon III detector in electron counting mode and a Volta phase plate. EPU software (FEI) was used for automatic data collection. Data were collected in nine independent sessions to give a total of 5,737 movies. Each micrograph was collected as 75 movie frames at a dose rate of 0.5 e^-/pixel/sec (0.4 $e^-/Å^2$ per frame) for 60 s, with a total accumulated dose of ~30 $e^-/Å^2$. The magnification used was 75,000×, yielding a pixel size of 1.06 Å/pixel.

Data processing and model building. RELION-2.1 was used for all data processing[31] unless otherwise specified. Since data were pooled from nine independent sessions we provide here the general strategy for data collection and processing, while precise particle numbers for a representative dataset are presented in Extended Data Fig. 3. Overall, drift, beam induced motion and dose weighting were corrected with MotionCor2[32] using 5 × 5 patches. CTF fitting and phase shift estimation were performed using Gctf v.0.1.06[33], which yielded the characteristic pattern of phase shift accumulation over time for each position. Generally, 40 images were taken at each Volta phase plate position. Auto-picking was performed with a Gaussian blob as a template[34] which readily resulted in optimal particle picking. Particles were extracted in a box of 150 pixels (159 Å) and inputted into a one or two reference-free 2D classification (if the majority of 2D classes had non-recognizable or low-quality features, then the selected particles belonging to quality classes were taken to a second round of 2D classification). An ab initio model was generated using 10,000 particles with RELION 2.1[35] in the first data collection and used throughout. The resulting particles after 2D classification where then used for 3D classification in both three and four classes simultaneously in order to check for consistency in 3D classification and to generate models with different numbers of particles. The models with the best defined features were selected for refinement either on their own or together with a second class from the same 3D classification (if more than one quality model was present). The particles that reached the highest resolution after gold-standard resolution estimation were saved. Particles obtained in a similar fashion from the different sessions were then merged and refined together. During refinement, the low-pass filter effect of the Wiener filter in the regularised likelihood optimisation algorithm was relaxed through the use of a regularisation parameter ($T = 3$). This allowed the refinement algorithm to consider higher spatial frequencies in the alignment of the individual particles. Nevertheless, both half-reconstructions were kept completely separately, and the final resolution estimate (at the post-processing stage in RELION) was based on the standard Fourier shell correlation (FSC) between the two unfiltered half-reconstructions. The final model contained 730,118 particles and reached an overall resolution of 3.78 Å with side chains visible for most of the complex (Extended Data Figs. 1, 2). Local resolution estimates were calculated with Resmap[36] showing a core of the protein at ~3.5 Å resolution and an extracellular region of the receptor and βγ N termini at poorer resolution with the worst regions reaching ~5 Å (Extended Data Fig. 1). Signal subtraction of the DM micelle did not improve the quality of the map upon refinement.

Model building and refinement was carried out using the CCP-EM software suite[37]. The 5-HT$_{1B}$R-ergotamine crystal structure was used as a starting model (Protein Data Bank (PDB) accession 4IAR)[15] for receptor building. 5-HT$_{1B}$R was modelled from residue L45 to R385. Although density was present from Y38 and this region seems to adopt a similar conformation to the 5-HT$_{1B}$R crystal structure, the poor resolution in this region prompted us to leave it unmodelled. Residues R188 to V196 in the ECL2 and I339 to C344 in ICL3 were flexible with absent or very poor map density and were therefore, not modelled. For the same reason residues K241 to L304 forming the large 5-HT$_{1B}$R ICL3 loop were left unmodelled. Mini-G$_o$ was modelled from residue L5 to Y354 following native Gα_o numbering. Modifications in Gα_o to obtain mini-G$_o$ are as described[10] (Extended Data Fig. 7). Although β and γ subunits were modelled using the available crystal structures, poor density was found for both N termini, with the whole of the γ subunit having poor density. For this reason the worst regions of these subunits were modelled as poly-alanine. Initial manual model building was performed in Coot[38] following a jelly-body refinement in REFMAC5[39]. Donitriptan coordinates and library were created with JLigand[40] and manually fitted into the density using sphere real space refinement in Coot. Restraints were generated with ProSMART[41] in order to maintain structural features in regions of poorer density. *B* factors were reset to 40 Å^2 before refinement. The model then followed cycles of manual modifications in Coot and restraint refinement in REFMAC5. The final model achieved good geometry (Extended Data Table 1) with validation of model performed in Coot, Molprobity[42] and EMRinger[43]. The goodness of fit of the model to the map was carried out using Phenix[44] using a global model-vs-map FSC correlation (Extended Data Fig. 2). Overfitting in refinement was monitored throughout using FSC_{work}/FSC_{test}[45].

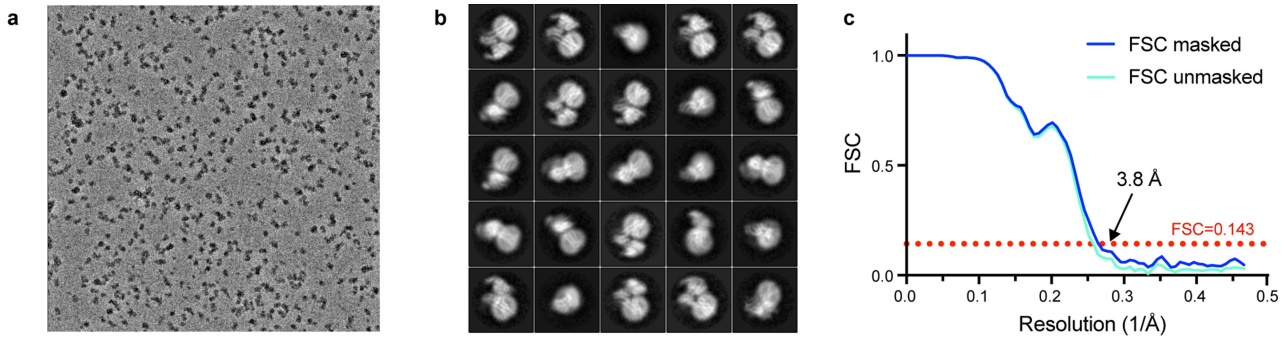
Note on limitations of the interpretation of density in the ligand binding pocket. The ligand binding pocket of 5-HT$_{1B}$R is occupied by a single molecule of the agonist donitriptan. Despite the resolution varying between 3.8 Å to 4.3 Å in this region, estimated from the local resolution map (Extended Data Fig. 1), the density allowed modelling of the position and orientation of donitriptan and the majority of the amino acid side chains in the pocket. However, the resolution limits the accuracy of the refined coordinates and care must be taken when analysing the precise details of any potential interactions. The best resolution is towards the centre of the membrane bilayer and resolution gets worse towards the extracellular surface of the receptor. The ligand has been modelled using real space refinement, taking into account the location of nearby residues as well as using a library of restraints with allowed conformations of donitriptan. The density allowed modelling of the position and orientation of the donitriptan molecule, with the indole group buried deep in the orthosteric binding pocket and the remainder of the ligand protruding towards the extracellular surface.

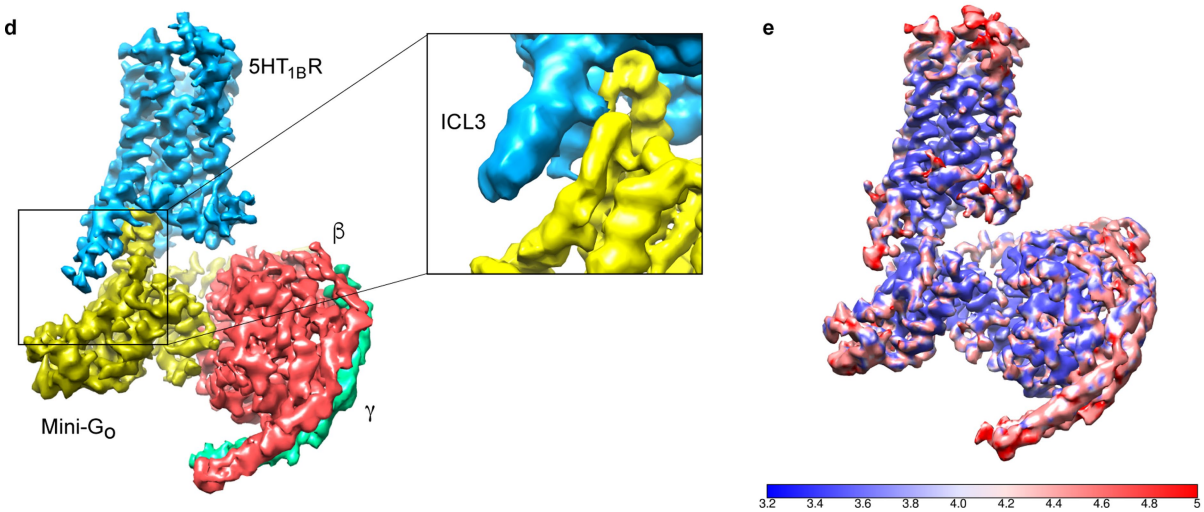
Clear interpretable density is found for large aromatic groups, while density is poorer for residues with smaller side chains such as Ser334$^{6.55}$ and Ser212$^{5.42}$. Tyr359$^{7.43}$, Phe330$^{6.51}$ and Phe331$^{6.52}$ are positioned around the indole group at the base of the pocket and Phe351$^{7.35}$ interacts with the donitriptan aromatic moiety at the most extracellular region. Met337$^{6.58}$ is located in a region of poor density and has been modelled so that is oriented away from the pocket and interacting with the aromatic group of donitriptan. This was concluded based on interpretation of maps with different sharpening levels, but its rotamer cannot be assigned with confidence. The orientation of the primary amine on the serotonin moiety in donitriptan and the adjacent side chain of Asp129$^{3.32}$ cannot be confidently assigned owing to poor density. However, Asp129$^{3.32}$ is absolutely conserved in all the human serotonin GPCRs and forms a hydrogen bond with ergotamine in the high-resolution crystal structure of 5-HT$_{1B}$R. We have therefore modelled Asp129$^{3.32}$ in a similar rotamer to make a potential hydrogen bond with this primary amine in donitriptan, despite the lack of density for both the primary amine and the carboxyl group of Asp129$^{3.32}$.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

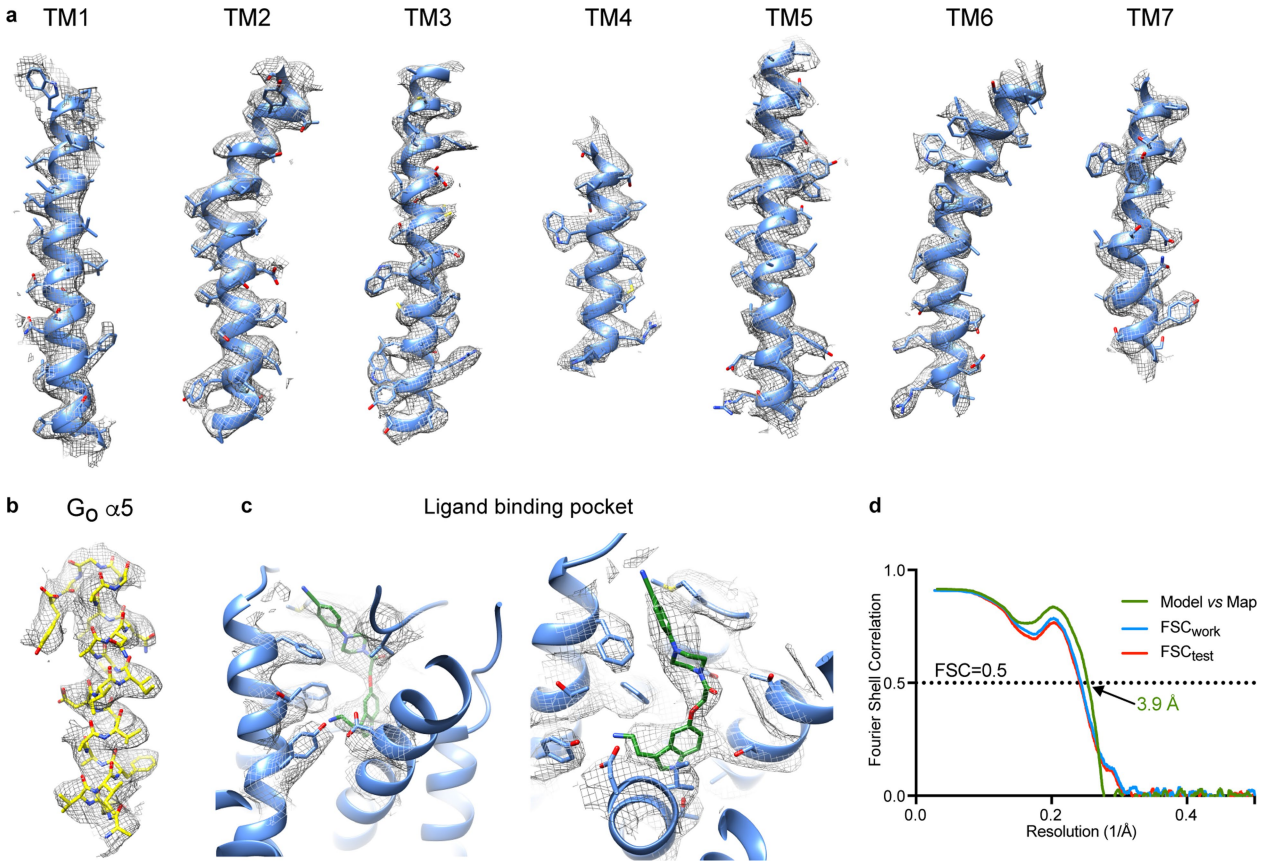
Data availability. All data generated or analysed during this study are included in this published article and its Supplementary Information. The cryo-EM density map has been deposited in the Electron Microscopy Data Bank under accession code EMD-4358 and the coordinates have been deposited in the Protein Data Bank under accession number 6G79.

31. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, https://doi.org/10.7554/eLife.18722 (2016).
32. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
33. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
34. Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr D* **73**, 496–502 (2017).
35. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
36. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
37. Burnley, T., Palmer, C. M. & Winn, M. Recent developments in the CCP-EM software suite. *Acta Crystallogr D* **73**, 469–477 (2017).
38. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D.* **60**, 2126–2132 (2004).
39. Murshudov, G. N. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D.* **67**, 355–367 (2011).
40. Lebedev, A. A. et al. JLigand: a graphical tool for the CCP4 template-restraint library. *Acta Crystallogr. D.* **68**, 431–440 (2012).
41. Nicholls, R. A., Long, F. & Murshudov, G. N. Low-resolution refinement tools in REFMAC5. *Acta Crystallogr. D.* **68**, 404–417 (2012).
42. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D.* **66**, 12–21 (2010).
43. Barad, B. A. et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).
44. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D.* **66**, 213–221 (2010).
45. Amunts, A. et al. Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**, 1485–1489 (2014).

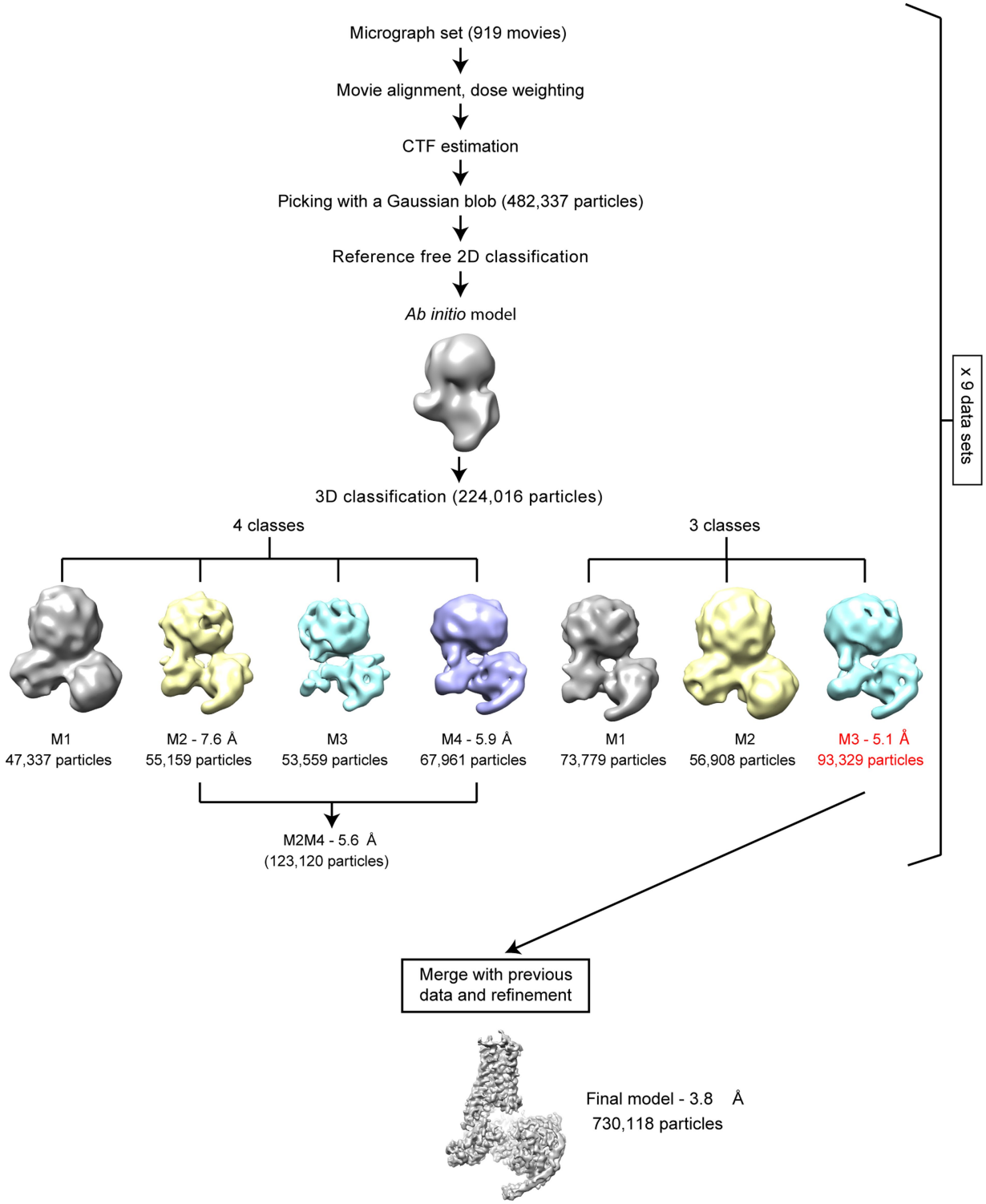




Extended Data Fig. 1 | Cryo-EM single particle reconstruction of the 5-HT_{1B}R–G_o complex structure. a, Representative micrograph (magnification 75,000×, defocus −0.6 μm) of the 5-HT_{1B}R–G_o complex collected using a Titan Krios with the Falcon III detector and Volta phase plate. **b**, Representative 2D class averages of the 5-HT_{1B}R–G_o complex. **c**, FSC curve of the final reconstruction showing an overall resolution of 3.8 Å using the gold-standard FSC of 0.143. Both masked and unmasked FSC curves are shown to highlight the lack of masking artefacts. **d**, Final reconstruction coloured by subunit. Inset shows a magnified view of the weak density for ICL3. The magnified region corresponds to a map sharpened with $B = -50$ to remove noise from lower density levels. **e**, Local resolution estimation of the 5-HT_{1B}R map as calculated by Resmap.



Extended Data Fig. 2 | Cryo-EM map quality and model validation. **a**, Transmembrane helices of 5-HT_{1B}R. **b**, The α5-helix of G_o. **c**, Donitriptan and the neighbouring side chains in the orthosteric binding site. **d**, FSC of the refined model versus the map (green curve) and FSC_{work}/FSC_{test} validation curves (blue and red curves, respectively).

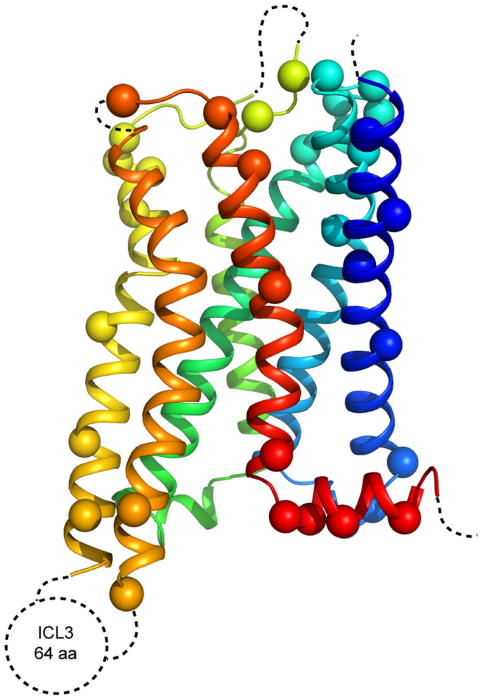


Extended Data Fig. 3 | Flow chart of data processing. Micrographs were collected during nine sessions on the Titan Krios (either 24 h or 48 h) and each session was processed independently. The number of images and particles from one 48-h session is indicated on the flowchart as a guide. At the bottom of the figure, the final number of particles is shown. Each dataset was corrected separately for drift, beam-induced motion and radiation damage. After CTF estimation, particles were picked using a Gaussian blob and submitted to either one or two rounds of reference-free 2D classification (see Methods). A 3D classification was performed on the selected particles using an ab initio model generated from ten thousand particles. Classification was performed in parallel in three and four classes. The models with best features were refined on their own; if there were two classes of similar high quality, these were then re-refined together (the resolution of the models refers to the resolution after refinement and calculation of gold-standard FSC = 0.143). The set of particles that obtained the best map quality and resolution were saved and merged with the best particles from other datasets. A final model with 730,118 particles was refined and achieved a global resolution of 3.78 Å.

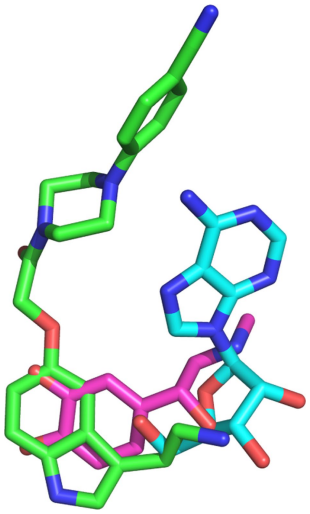
a

```
34   SAKDYIYQDSISLPWKVLLVMLLALITLATTLSNAFVIATVYRTRKLHTPANYLIASLAV   93
94   TDLLVSILVMPISTMYTVTGRWTLGQVVCDFWLSSDITCCTASIWHLCVIALDRYWAITD  153
154  AVEYSAKRTPKRAAVMIALVWVFSISISLPPFFWRQAKAEEEVSECVVNTDHILYTVYST  213
214  VGAFYFPTLLLIALYGRIYVEARSRILKQTPNRTGKRLTRAQLITDSPGSTSSVTSINSR  273
274  VPDVPSESGSPVYVNQVKVRVSDALLEKKKLMAARERKATKTLGIILGAFIVCWLPFFII  333
334  SLVMPICKDACWFHLAIFDFFTWLGYLNSLINPIIYTMSNEDFKQAFHKLIRFKCTS     390
```

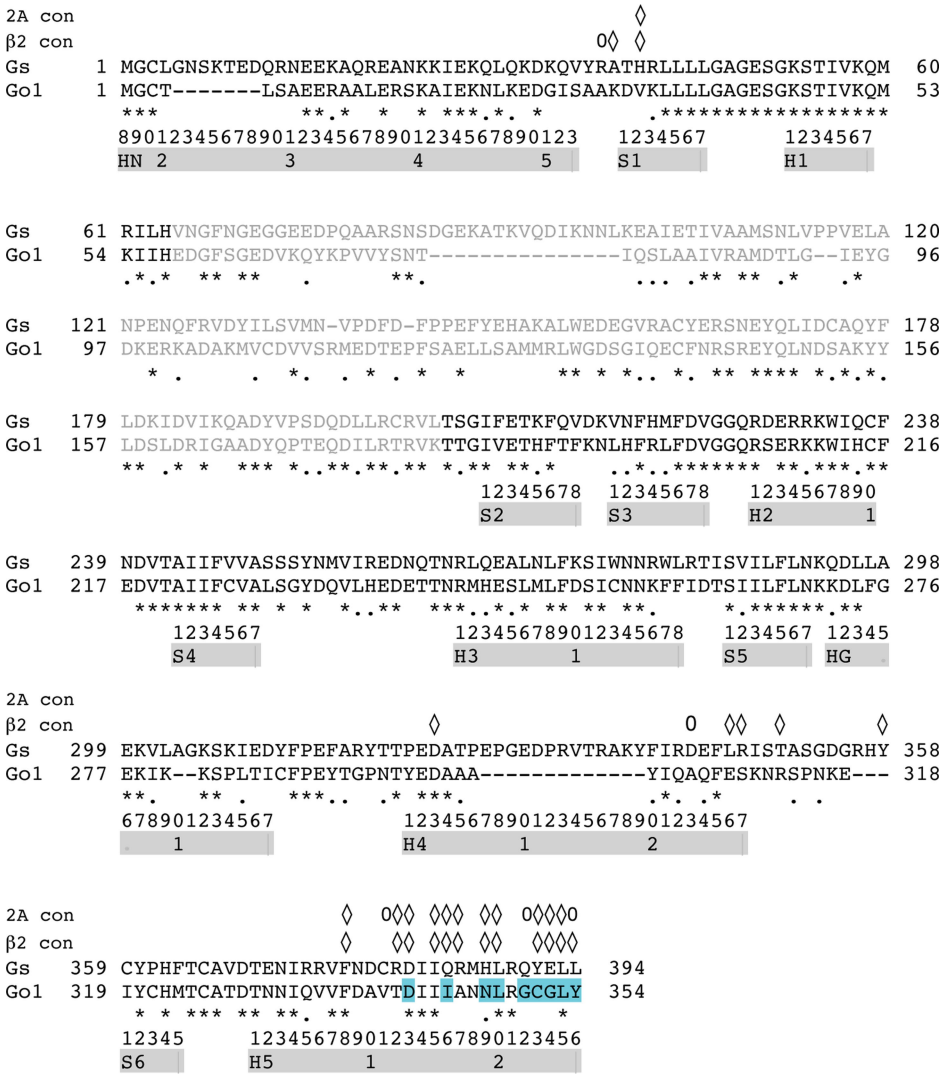
b



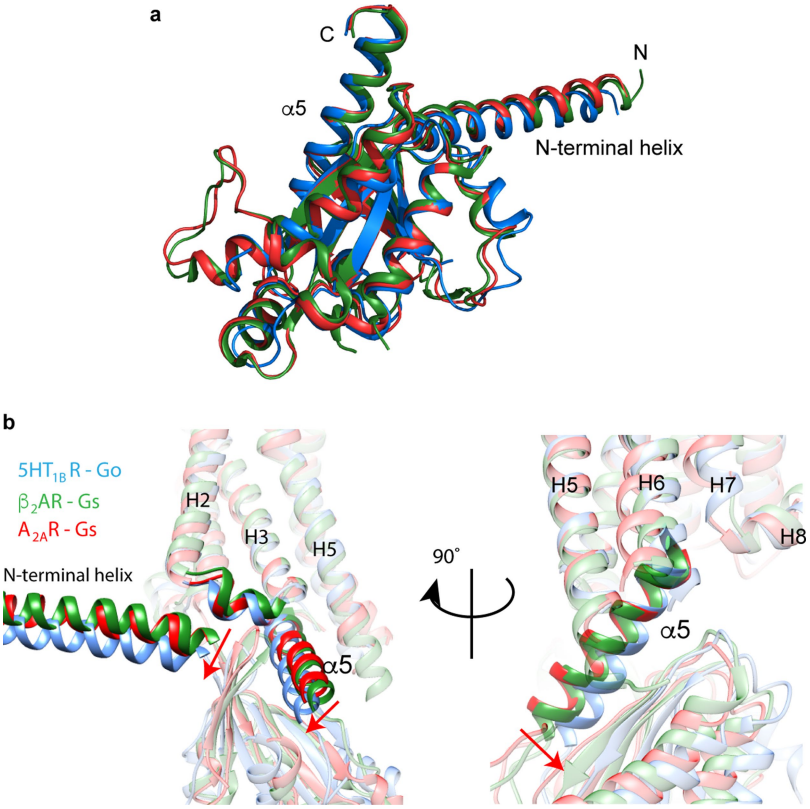
Extended Data Fig. 4 | Modelling quality of the 5-HT$_{1B}$R structure. **a**, Amino acid sequence of the 5-HT$_{1B}$R construct used for the cryo-EM structure determination. Residues are coloured according to how they have been modelled. Black, good density allows the side chain to be modelled; red, limited density for the side chain, therefore the side chain has been truncated to C_β; blue, no density observed and therefore the residue was not modelled. Regions highlighted in grey represent the transmembrane α-helices, and amphipathic helix 8 is highlighted in yellow. **b**, Model of 5-HT$_{1B}$R showing the C_α positions of amino acid residues with poor density (spheres) and unmodelled regions (dotted lines).



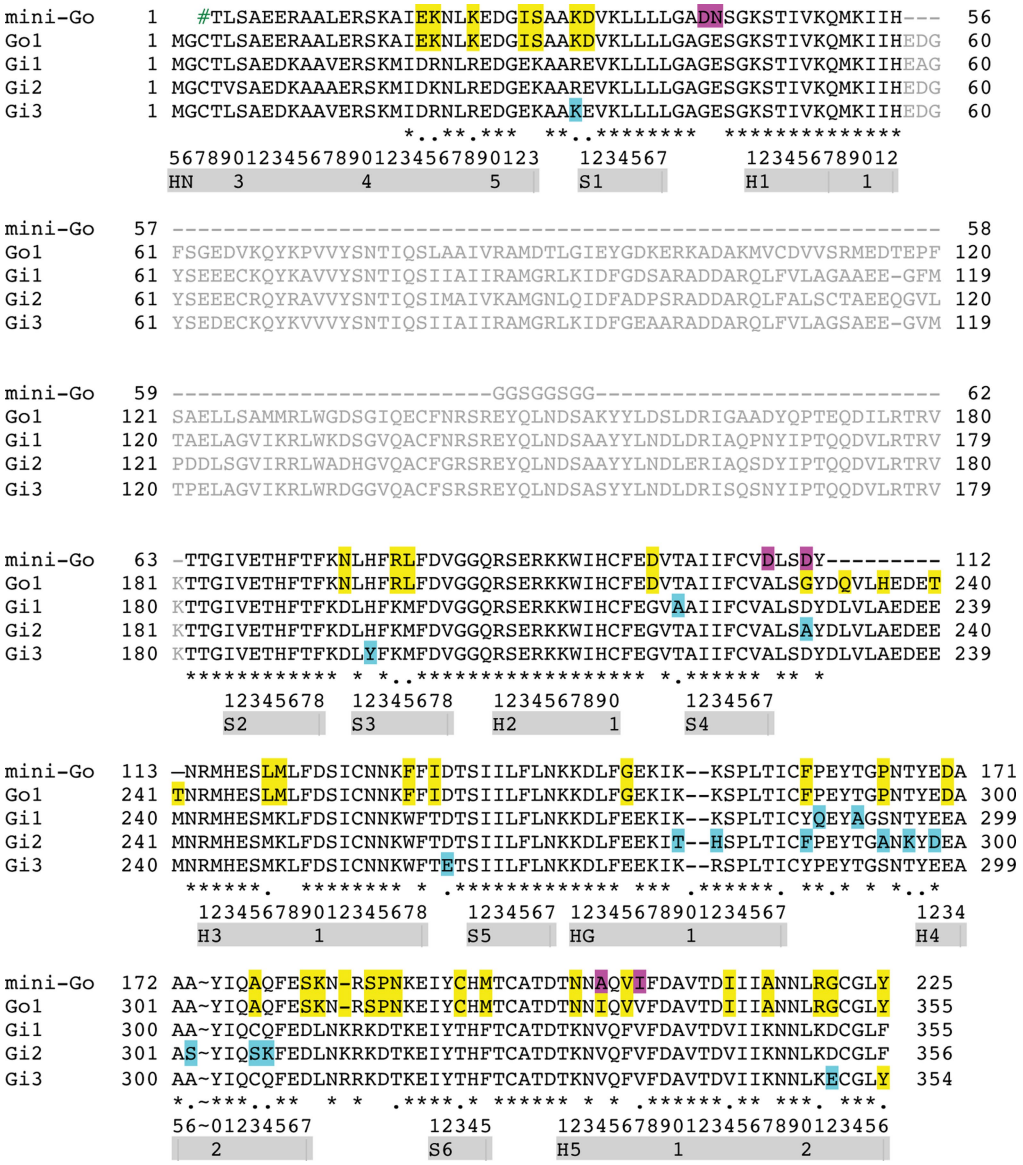
Extended Data Fig. 5 | Superposition of donitriptan, adrenaline and adenosine bound to their respective receptors. 5-HT_{1B}R, β_2AR[3] and A_{2A}R[1] were superimposed (using Pymol) over the whole of the receptor. Green, donitriptan; pink, adrenaline; blue, adenosine.



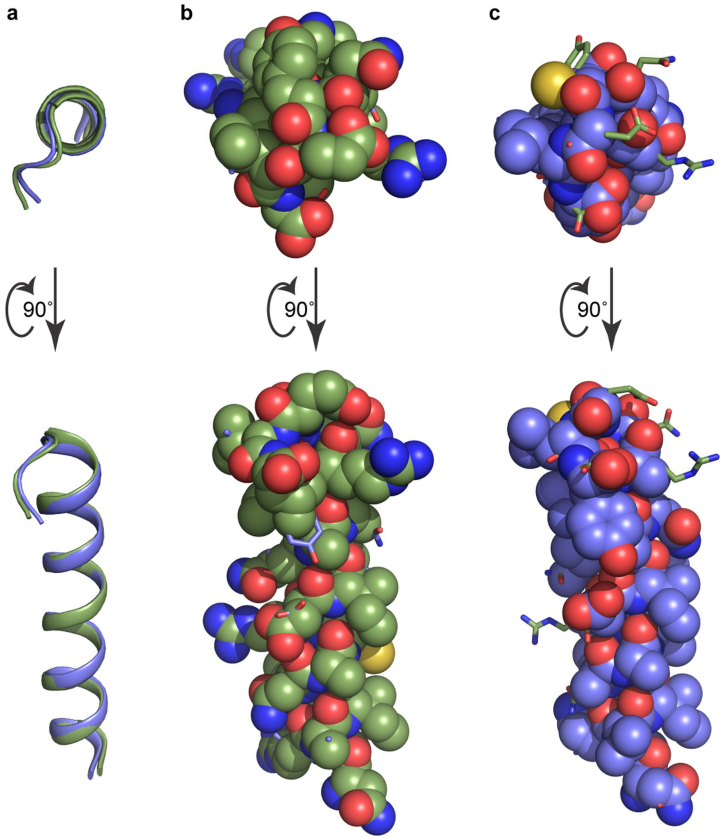
Extended Data Fig. 6 | Comparison of the amino acid sequences of the α-subunits of G_o and G_s. Diamonds above the sequences indicate amino acid residues in $G\alpha_s$ in which the side chains make atomic contacts to residues in β_2AR (β2 con) or A_{2A}R (2A con). Ovals indicate amino acid residues in $G\alpha_s$ in which only the main chain atoms make contacts with the receptor. Secondary structural elements are indicated as grey bars with positions numbered according to the CGN numbering system.



Extended Data Fig. 7 | Similarity of Gα structures and the difference poses of the α5-helices in $G\alpha_o$ and $G\alpha_s$ coupled to receptors. a, The structures of the α-subunits in complex with 5-HT_{1B}R, β_2AR[3] and A_{2A}R[1] were superimposed over the whole of their sequence in Pymol. Blue, $G\alpha_o$ coupled to 5-HT_{1B}R; green, $G\alpha_s$ coupled to A_{2A}R; $G\alpha_s$ coupled to β_2AR. **b**, 5-HT_{1B}R (blue), β_2AR[3] (green) and A_{2A}R[1] (red) were superimposed based on H3, H5 and H6. Two different views are shown with the red arrows indicating differences in orientation of $G\alpha_s$ and $G\alpha_o$.



Extended Data Fig. 8 | Alignment of the amino acid sequences of G_o and G_i α-subunits. Residues in grey correspond to the α-helical region that does not make contact with GPCRs and was deleted during the construction of mini-G_o. Secondary structural elements are depicted as grey bars with the CGN numbers shown to aid comparisons. Amino acids are highlighted as follows: pink, stabilizing residues required to generate mini-G_o; yellow; residues in $G\alpha_o$ that are different from residues conserved in all three $G\alpha_i$ sequences; blue, residues that are non-conserved in $G\alpha_i$ sequences. #, the affinity tag on mini-G_o used for purification (MGHHHHHHENLYFQG).



Extended Data Fig. 9 | Comparison of the α5-helices of G_s and G_o. The α5 helices in the cryo-EM structures of $A_{2A}R$–G_s (carbon, green) and 5-$HT_{1B}R$–G_o (carbon, light blue) were aligned (in Pymol) along their whole sequence and displayed in different poses: cartoon depiction (**a**); G_s (green spheres), G_o (blue sticks) (**b**); G_o (blue spheres), G_s (green sticks) (**c**).

Extended Data Table 1 | Data collection and refinement statistics

	5-HT_{1B} - $\text{MiniG}_{O}\beta\gamma$ (EMDB-4358) (PDB 6G79)
Data collection and processing	
Magnification	75,000x
Voltage (kV)	300
Electron exposure (e–/Å^2)	30
Defocus range (μm)	-0.3 to -1.0
Pixel size (Å)	1.06
Symmetry imposed	C1
Initial particle images[a] (no.)	1,249,822
Final particle images (no.)	730,118
Map resolution (Å)	3.78
FSC threshold	0.143
Map resolution range[b] (Å)	~3.4 to ~4.6
Refinement	
Initial model used (PDB code)	5G53, 3SN6
Model resolution[c] (Å)	3.9
FSC threshold	0.5
Map sharpening *B* factor (Å^2)	-200
Model composition	
Non-hydrogen atoms	6053
Protein residues	6023
Ligands	30
B factors (Å^2)	
Protein	97
Ligand	108
R.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	1.02
Validation	
MolProbity score	1.07
Clashscore	0.61
Poor rotamers (%)	0.56
EMRinger score	2.34
Ramachandran plot	
Favored (%)	94.64
Allowed (%)	4.88
Disallowed (%)	0.48

[a]After 2D classification.

[b]Local resolution range.

[c]Resolution at which FSC between map and model is 0.5.

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

BLOG Personal stories and careers counsel **blogs.nature.com/naturejobs**

TWITTER Links to career resources and tips **go.nature.com/2to9twc**

WILSON PANG

Scientists such as Yue Wan are reaping the rewards of international experience.

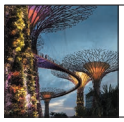
LAB LIFE

Full circle

Researchers who return home to East Asia aim to maintain networks and global standing.

BY VIRGINIA GEWIN

Many junior scientists from East Asia do their degree or post-PhD training abroad, but return home to develop their careers. The booming economies of several East Asian regions offer significant opportunities for new principal investigators. *Nature* spoke to four researchers who have recently returned to East Asia about the opportunities and challenges they face as they launch labs, seek talented lab members and forge and maintain collaborations.



HUBS OF ASIAN SCIENCE
A *Nature* collection
nature.com/collections/asianhubs

YUE WAN
Be adaptable

Structural genomicist, Genome Institute of Singapore

When I returned to Singapore in 2013 to start my own lab, I was excited to reacquaint myself with my Asian culture, especially as it becomes less conservative due to globalization. I had spent the previous decade in the United States, after receiving a scholarship from the Singaporean government's Agency for Science, Technology and Research (A*STAR) to develop well-trained PhDs to support Singapore's budding biomedical industry. Singapore is a small republic that built itself from almost nothing, with no natural resources, to a first-world city-state that prides itself on excellence, perseverance and resilience.

After earning my bachelor's degree in cell biology at the University of California, San Diego, I ended up at Stanford University in California for my PhD, where we developed a new protocol for conducting genome-wide measurements of RNA structure in yeast (M. Kertesz *et al. Nature* **467**, 103–107; 2010). We followed that work with studies of RNA-folding stability in yeast, and on the variation in RNA structure across the human transcriptome, the collection of gene readouts in a cell (Y. Wan *et al. Nature* **505**, 706–709; 2014).

Most A*STAR fellows return as postdocs, ▶

but in 2013, Singapore launched a fellowship programme that offered newly returned PhD holders enough money to start their own labs straight away. That same year, I became the first fellow at the Genomics Institute of Singapore. All of a sudden I had to work out how to hire people, create a budget and get orders in. It was a steep learning curve.

It was also an oddly isolating position. Most of my colleagues, and even I, didn't know what being an independent fellow meant. Postdocs didn't hang out with you, and neither did principal investigators (PIs). Former mentors advised me to make myself visible by speaking up at every staff meeting. That is hard, especially as a female scientist at an institute with few female PIs. But I realized that I had the money and independence to do what I wanted, so I did.

Singaporean research tends to be more applied and translational than US science — geared towards areas of strategic importance to society. One challenge is that priorities for funding can shift every 5–10 years to shape economic growth, so researchers must be adaptable.

In 2016, I received an A*STAR investigatorship, which gave me 6 million Singapore dollars (US$4.4 million) for 6 years and allowed me to take a senior-scientist post. I also earned several international fellowships, including a Branco Weiss Fellowship from the Swiss Federal Institute of Technology (ETH) in Zurich, a L'Oreal–UNESCO (United Nations Educational, Scientific and Cultural Organization) fellowship for women in science and an EMBO Young Investigator Award, which helped me to collaborate with scientists worldwide.

I've had my lab for five years, and have had two children during that time. Singapore currently offers four months' paid maternity leave for most citizens, but I didn't take it either time. I worried that my lab would tank if I were away for four months. I took three weeks off after each birth. For the following two months, I worked two hours a day to help my staff members to stay on course. Then I returned to work full time.

The Asian family structure is close knit. My parents live close by and are happy to take care of their grandchildren. Hired help is more affordable here than in the United States. Without those support structures — including my husband, who splits his time between his labs at the Genome Institute of Singapore and Nanyang Technological University — there would be no way for me to put in the hours necessary to run my lab.

NETHIA KUMARAN
Maximize visibility

Cancer biologist, University of Science Malaysia

Since returning to Malaysia after four years in Australia, I've noticed that people are more aware of the importance of science and engineering, and that more girls are encouraged to pursue science. I first became interested in cancer biology as an undergraduate in microbiology, when I learned that viruses could cause cancer. I was studying at the University of Science, Malaysia, and searched for a mentor, but there weren't many experts in that field in Malaysia at the time.

In 2005, after earning my bachelor's degree, I moved to the United Kingdom for a master's in oncology at the University of Nottingham. The Malaysian government offered full scholarships for people doing a PhD in or outside Malaysia. Because there were so few cancer-biology experts in Malaysia, I went to the University of Sydney, Australia, where I earned a PhD in 2012, studying the cell-death pathways of cancer.

"Global collaboration is not an easy process — sometimes I don't even get answers to e-mails I send to researchers overseas."

I had two excellent mentors in Australia. They helped me to better articulate my findings, learn to collaborate and to speak in public at international conferences. I focused on building my network and collaborations.

I wanted to do a postdoc to get more training and experience, but Malaysia requested that their funded scholars come back as soon as we finished our PhDs. It was a big jump to being a PI when I chose to return to the University of Science, Malaysia, in 2013. Located on the beautiful island of Penang, the university sits in one of the country's more culturally liberal areas.

My mentors in Sydney had taught me that getting a grant isn't just about the proposed research details. Success is also about clearly telling funders why this work is important, how it fits into the bigger picture and why they should give me the money to do it. Although that was helpful, I still needed mentors in Malaysia to help me navigate our grant system. The culture and academic protocols can be very different at home. It was a painful learning process, but I secured grants from the university and one from the Ministry of Higher Education.

In my five years as a PI, research funding in Malaysia has been cut and grants have become increasingly competitive, but my lab has two PhD students and two master's students. Cancer biology is still a small community in Malaysia, although I do have a few collaborators here. I also apply for international grants, but that requires international collaborators. And global collaboration is not an easy process — sometimes I don't even get answers to e-mails I send to researchers overseas.

Right now, I'm establishing my visibility. Because I can't go to international conferences every year due to a lack of funding, I organized an international conference in Malaysia in 2017. Winning a L'Oreal–UNESCO fellowship in 2016, after applying for four years in a row, also really helped to improve my visibility.

Nethia Kumaran

NETHIA KUMARAN

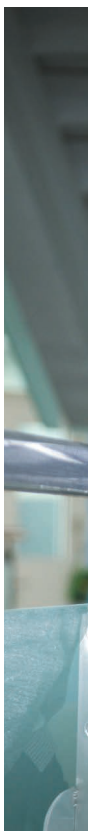
Having said that, I would have to think hard about leaving Malaysia to take a job in another country. The jobs here are permanent, whereas it seems like there are more temporary positions abroad. That scares me a bit. Ideally, I'll go on sabbatical overseas, pick up new skills and come back home.

MYUNGEUN SEO
Stay competitive

Polymer chemist, Korea Advanced Institute of Science and Technology

In 2012, I co-authored a paper in the journal *Science* with my postdoctoral adviser Marc Hillmyer, a chemist at the University of Minnesota in Minneapolis (M. Seo *et al. Science* **336**, 1422–1425; 2012). It took two years to complete, but he gave me the time and space to put together a whole story. A two-author paper in a high-impact journal is rare these days given the pressure to publish, but I believe that it helped me to secure my current job at Korea Advanced Institute of Science and Technology (KAIST), where I also earned my master's and PhD.

Since my appointment in 2013, I'm happy to be back in the science-oriented city of Daejeon,



which has more than 30 national laboratories as well as industrial research centres. As a chemistry student at KAIST, I visited laboratories in Japan and the United States while researching the self-assembly of small molecules into nanostructures. In 2007, during my PhD, I spent six months as a visiting researcher at the University of California, Santa Barbara, where I learnt a new type of polymerization — one that I still use today, and which has had a huge impact on my research path. I published a paper (M. Seo *et al. Macromolecules* **41**, 6413–6418; 2008) that has been cited 100 times so far. These experiences encouraged me to apply for several postdocs abroad, which is how I ended up in Marc's lab in 2009.

After four years, I returned to South Korea, and the on-site interview for my job at KAIST was one of the toughest days of my life. The institute competes with other universities for scientists who will be highly visible in their fields. But the country is experiencing a big wave of retirements right now, so there are good job opportunities.

My start-up package let me purchase a lab bench, fume hood and some crucial instruments to fill my empty space in a new building. I started without a postdoc or technician because it is difficult to find good candidates here. I have 13 students, 4 of whom are female, and 1 postdoc, from India. KAIST successfully increased the number of foreign students and foreign and female faculty members to 10% of the total school population and now aims to increase those numbers by another 10%. I try to give my students the time they need to put a research story together, as Marc did for me. Three of the papers I have published since starting my lab have had a student as the only co-author.

The grant funding situation fluctuates year to year — and is highly competitive, with a roughly 10% success rate. Despite this, I recently learnt that in addition to a personal grant to continue my basic research, a team that I am part of has won a 7-year grant of US$10 million to explore the chemical architecture of self-assembled molecules.

I'll probably apply for tenure in 2020. Since KAIST's graduate school of nanoscience and technology launched in 2008, only one person has sought, and ultimately achieved, tenure. Publishing a number of papers is always helpful, but the content is considered more important, as is which journal you publish in. International visibility is also key to securing tenure. In March, with US colleagues, I organized symposia at the annual American Chemical Society meeting in New Orleans, Louisiana, that brought together world leaders studying how complex polymers self-assemble.

KAIST KI

Myungeun Seo

KIM HEI-MAN CHOW
Stay connected

Neuroscientist at Hong Kong University of Science and Technology

In 2013, Karl Herrup, a neuroscientist at Hong Kong University of Science and Technology, advertised for someone with expertise in Wnt signalling, a crucial pathway that regulates the fate of cells. He hired me, even though I was a cancer biologist with no background in neuroscience, in hope that a fresh perspective would prove valuable.

I'd studied Wnt signalling as a molecular-biology PhD student at the University of Hong Kong. During my third year, I presented my research identifying a potential drug target that could regulate Wnt signalling at the American Association for Cancer Research meeting in Washington DC.

"The bond is stronger when you can talk in person and discuss current research together."

After my presentation, I received two invitations to pursue a postdoc in the United States.

I chose to go to Cornell University in Ithaca, New York, to work with biomedical engineer Xiling Shen in 2010, but I had never used a computational-biology or systems-biology approach. My experience at Cornell rewired my approach to biomedical research.

One reason that clinical trials can fail is that we tend to look at only one thing at a time. With systems biology, we first look at the whole picture and screen whether a gene or protein is the major cause of disease. Then we use molecular-biology tools to focus the research. Six months after arriving at Cornell, however, I had to return home to care for my sick mother. I couldn't devote myself to full-time research, so I worked for a medical editor for a year while my mother recovered.

My first year in Karl's lab was frustrating, because I had to learn about the brain from scratch, but I decided to seize this chance to expand my skills and launch a second career applying a systems-biology approach to studying Alzheimer's disease.

In my experience, students and postdocs in the United States tend to be freer to explore ideas that are unrelated to the PI's work. In Hong Kong, most students do research that has already been laid out by their supervisor. Both ways have pros and cons. Hong Kong attracts a lot of foreign professionals because it's very accessible to cities such as Tokyo, Taipei, Seoul and Singapore. It is also a cosmopolitan metropolis where old tradition blends with Western culture and post-modern trends — perfect for a foodie like me.

Funding opportunities are perhaps the biggest difference between the United States and Hong Kong. In the United States, there are federal and non-profit foundations that fund small projects or fellowships for postdocs; in Hong Kong, there is very little, and most early-career scientists are unaware of international opportunities.

I published my first neuroscience paper within 12 months of joining Karl's lab, and have since secured grants from the US Alzheimer's Association as well as from the Institute for Advanced Study.

I have been a research assistant professor for the past three years. It is a 'grey' position between postdoc and fully independent assistant professor. I can apply for grants and conduct my project independently, but I'm still affiliated with Karl's lab.

I'm also building my profile in the field through other fellowships, including one with the World Economic Forum, to explore other collaborations as I prepare to start my own lab, which I hope to do next year.

To maintain international relationships, I make myself available for Skype calls in the evenings and on weekends. And, crucially, I see colleagues such as Shen in person whenever I can. The bond is stronger when you can talk in person and discuss current research together.

My goal is to get a couple of high-impact publications before I start looking for my next job. Hong Kong is a small place, and space is limited owing in part to an influx of researchers from mainland China, so I'm open-minded about eventually taking a position elsewhere. ■

INTERVIEWS BY VIRGINIA GEWIN

These interviews have been edited for clarity and length.

TRAUMAHEAD

Preservation strategy.

BY JEREMY SZAL

I'm crunching over the bodies of the fallen when I see him: the warrior with the head swollen with nanomemories.

My skin is raw and bleeding where my black CombatSkin has been shredded open by human gunfire. I can smell my sour wounds. My weapons harness chafes my neck and I can't stop shivering against the icy wind. Nothing burns like the cold. But self-pity must wait. I grit my teeth and trudge over to the fallen warrior. I press my eight-fingered hand to the nape of his blood-stained neck, where his shattered helmet has exposed his bulging nanomemories. The nanofibres in my skin tighten, red lattice-work standing stark as the warrior's neck peels from his body in luminescent cubes. His vibrating nanomemory cells float to my outstretched hand, disintegrating on contact and merging with my nanofibre matrix.

It's like having a bucket of ice-cold water drip down my spine as the memories spear into my head. We were told to store only the highlights. No unnecessary clutter, no memories that others might share. I watch years of this warrior's life play out in fragments: his first time seeing the sprawling cityscape we built on this frozen, tidally locked moon. Meeting his wife the night we gathered to watch the solar flare. The day he signed up to fight the human invaders, after they broke our peace treaty.

The birth of his daughter.

My throat is locked tight as I stand. My neck and shoulders prickle with sweat, scarred flesh expanding as the warrior's nanomemories are rewritten to my biometrics. Fusing with my body, my consciousness on a molecular level. But the only memories I possess of Asher, my daughter, are my own. Her head resting on my chest as a child, wanting to listen to her father's heartbeat. Her insistence on joining the armed forces when she became old enough. Wearing that stubborn, determined look that reminds me so much of myself at her age. But the memories are grey, distorted. They don't have the crystalline clarity of nanomemory infusion. They're inscribed in pencil, not ink.

I continue along the frozen battlefield in search of nanomemories, hoping I will find hers. Ash and snow melt on my shoulders and soak into my CombatSkin. Before long, I've scavenged the nanomemories of 50 of our slaughtered warriors. Childhood upbringings, deaths, loves, disappointments. Scientific and technological breakthroughs. Seeing friends and family blown to twitching bundles of flesh by human artillery. Brave soldiers donning their CombatSkins and thermal helmets and departing for the battlefield, knowing they will never return home. My numb muscles feel like flayed meat, my spinal column stretching as my body processes the constant stream of narratives.

Some guessed the peace treaty between our race and the humans would not last. Our existence interrupted their colonization plans for this moon, after all. And if there's one thing we know about humans: they do not like to share. Our Jhulivaan civilization predates theirs by 15 millennia, but they have the numbers, the determination, the sheer brute force to destroy us. It's not enough to take our home; they want to drive us to extinction. Pretend that we never existed, that they are the only race in the Universe. Despite their small physical stature, stout limbs and excessive body hair, we're so similar in appearance that perhaps they took our existence as an insult.

Only, we were prepared for this. We've archived our memories, our histories, fragments of our civilization, in nanomemories to be retrieved upon death. But memories are based on emotion. AIs process data and raw statistics. Emotional states are unreadable to them. The only thing that can store conscious experiences is another conscious being.

I allowed the nanofibre matrix to be installed atop my spinal column. It houses these fragmented memories, laced with love, hatred, acceptance, rage, fear — and the events that led up to them. I rub the fuzz along my swollen neck, not quite believing that I'm becoming pregnant with the ghosts of an entire civilization.

But no matter how many piles of mutilated corpses I sift through, no matter how many lifetimes and personalities I transfer into me, I cannot find my daughter. I cannot find her memories of me and our brief 90 lunar cycles together before the war. My weariness vanishes, my chest heaving as I run from body to body, hoping I will find her so she can live on inside me. The sea of dead faces gazes up at me with eyes like bullet holes. But none is hers. The howling wind lashes at my maimed face. I want to sink to my knees and punch the ice until my hands are blocks of bloody, frozen flesh.

I swallow, look up. There's a human warship in orbit. They'll be here soon,

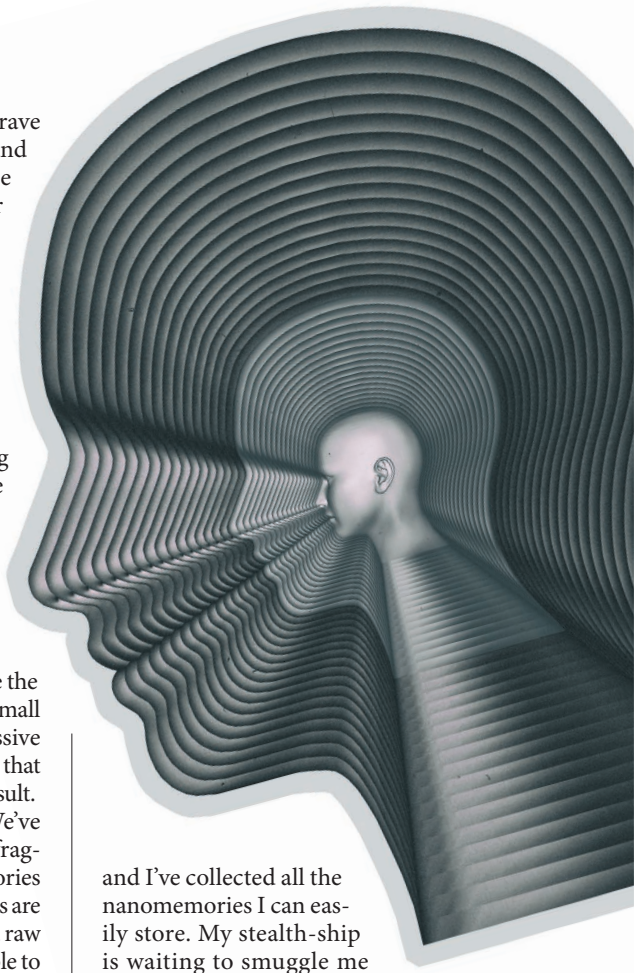


ILLUSTRATION BY JACEY

and I've collected all the nanomemories I can easily store. My stealth-ship is waiting to smuggle me off-world to the remaining survivors, where we'll stash our pool of nanomemories in a place inaccessible to humans. Let the haughty bastards think they've destroyed all trace of us. And one day, maybe a spacefaring explorer will uncover our time capsule and share who we once were with the Universe.

I exhale slowly and take one final glance at the smoking remains of the cityscape before crunching away over the ice. My nape and shoulders have swollen to twice their normal size. But there's a gaping hole in there, like a wound. I think of my daughter's face, her laughter, her anger, her stubborn bravery and outbursts, and all the things that made her mine. I know that one day, when I die or become senile, the memories will fade to smoke in my hands. But until then, as long as I can hold her, she'll be with me.

I press my hand over my heart as I stumble away from my home, pretending it's hers. ■

Jeremy Szal *writes about galactic nightmares, little traumas and broken heroes looking for hope in dark worlds from Sydney, Australia. Get in touch at jeremyszal.com or @jeremyszal*

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm